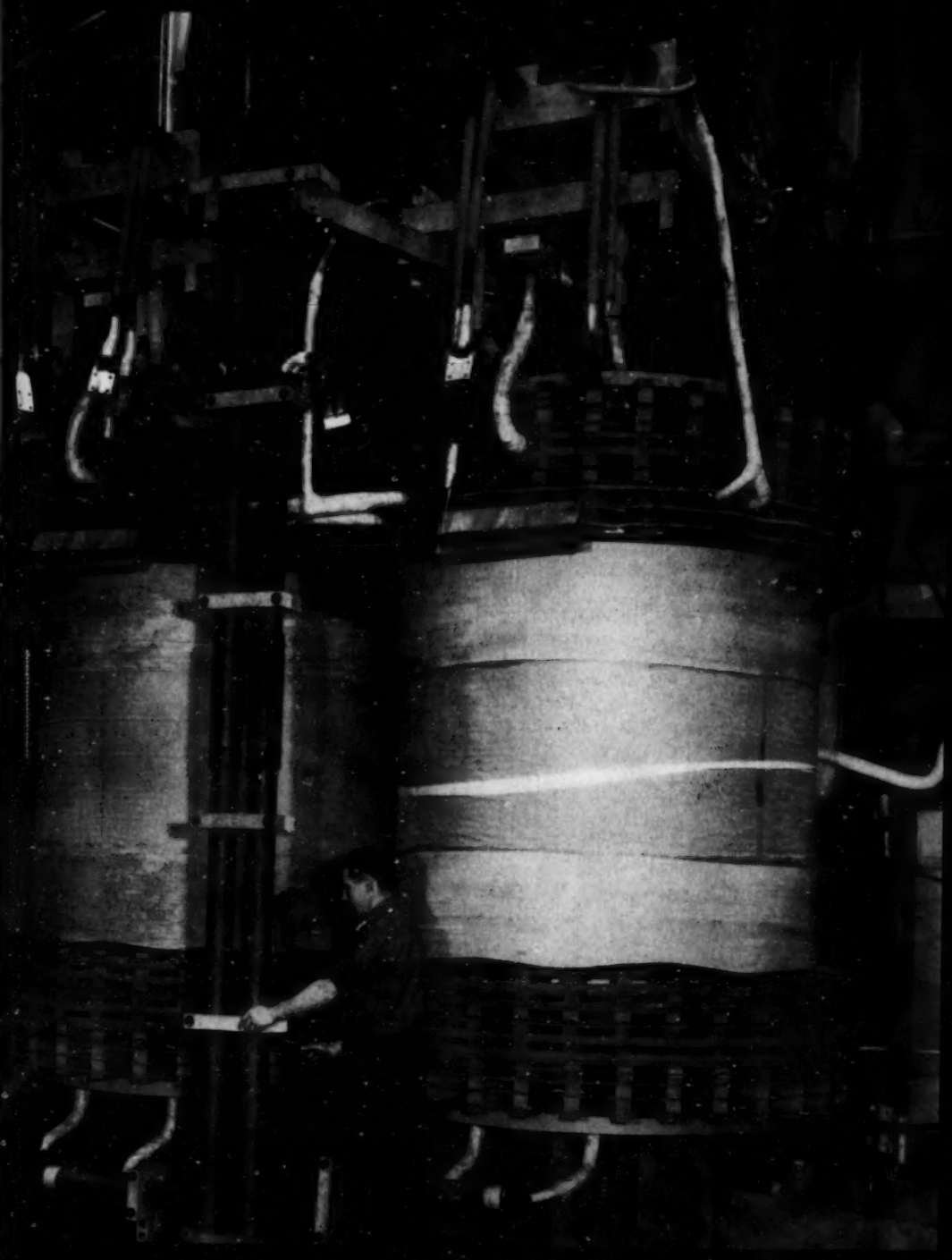
JOURNAL OF APPLIED PHYSICS



JOURNAL OF APPLIED PHYSICS

Vol. 24, No. 4 APRIL, 1953

In This Issue

		Page
Statistical Criteria for the Detection of Pulsed Carriers in Noise. I	DAVID MIDDLETON	371
Statistical Criteria for the Detection of Pulsed Carriers in Noise. II	DAVID MIDDLETON	379
Use of Tschebyscheff-Polynomial Operators in the Numerical Solution of Boundary-Value Problems	GEORGE SHORTLEY	392
Optimum Nonlinear Filters	L. A. ZADEH	396
Effects of Very High Pressures on Glass	P. W. BRIDGMAN AND I. ŠIMON	405
Susceptance of a Circular Obstacle to an Incident Dominant Circular-Electric Wave	L. S. SHEINGOLD	414
Bulk Viscosity of the System Polystyrene—Diethyl Benzene	F. BUECHE	423
Positive and Negative Ions in Cathode-Ray Tubes	C. H. BACHMAN, G. L. HALL, AND P. A. SILBERG	427
Effective Length of an Electrostatic Analyzer	JAMES E. MONAHAN	434
Electrical Phenomena in Adhesion. I. Electron Atmospheres in Dielectrics	SELBY M. SKINNER, ROBERT L. SAVAGE, AND JOHN E. RUTZLER, JR.	438
Effect of Turbulence on Heat Transfer in a Pile with a Liquid Reactant	E. N. PARKER	451
Retarding Potential Method for Measuring Electrical Conductivity of Oxide-Coated Cathodes	I. L. SPARKS AND H. R. PHILIPP	453
Engravement of Transient Stress Wave Particle Velocities	JOHN S. RINEHART AND JOHN PEARSON	462
Photoelastic Studies of Quenched Cylinders and Spheres	K. A. PARSONS	469
Work Functions of Gas-Coated Tungsten and Silver Surfaces	G. L. WEISSLER AND T. N. WILSON	472
On the Zeros of Polynomials and the Degree of Stability of Linear Systems	J. FRANK KOENIG	476
Approximate Elastic Spectrum of Acoustic Waves in AgCl from X-Ray Scattering	H. COLE	482
Limiting Negative Pressure of Mercury in Pyrex Glass	LYMAN J. BRIGGS	488
On Unsteady Flow of Compressible Viscous Fluids	G. LUDFORD, H. POLACHEK, AND R. J. SEEGER	490
Letters to the Editor:		
Simple Method for Measuring Interface Impedance	A. VAN DER ZIEL	496
Potential of Two Current Point Sources in a Homogeneous Conducting Prolate Spheroid	JAMES R. WAIT	496
Method for Increasing the Electrical Resistivity of Insulators Under Ionizing Radiation	JOHN H. COLEMAN AND DAVID BOHM	497
Thermoelectric Effects in Molybdenum Disilicide	MARTIN J. ARVIN	498
Permanent Magnet Properties of Cobalt-Platinum Alloys	D. L. MARTIN AND A. H. GEISLER	498
Comments on "A Note on Critical Reflections of Elastic Waves at Free Surfaces"	R. D. MINDLIN AND T. KANE	498
Comment on the Preceding Letter by R. D. Mindlin and T. Kane	W. S. JARDETZKY	499
Diffusion of D_2 from D_2O through Steel	FRANCIS J. NORTON	499
Solubility Determination of Some Elemental Gases in Pyrex at 1170°C	JOHN J. NAUGHTON	499
Announcements		500
Index to Advertisers		xxx

ELMER HUTCHISSON, *Editor*

EARLE C. GREGG, *Assistant Editor*

Board of Editors

RAY E. BOLZ
LYLE D. BORST
WALLACE J. ECKERT
ROBERT D. HEIDENREICH
LEONARD B. LOEB
W. JAMES LYONS
LOUIS A. PIPES
ROMAN SMOLUCHOWSKI
W. R. SMYTHE

Publication of the American Institute of Physics

OFFICERS

GEORGE R. HARRISON, *Chairman*
GEORGE B. PEGRAM, *Treasurer*

GOVERNING BOARD

J. W. BUCHTA
KARL K. DARROW
G. J. DIENES
S. A. GOUDSMIT
GEORGE R. HARRISON
FREDERICK V. HUNT
DEANE B. JUDD
HUGH S. KNOWLES
C. C. LAURITSEN
WILLIAM F. MEGGERS
PHILIP M. MORSE
BRIAN O'BRIEN
GEORGE B. PEGRAM
I. I. RABI
DUANE ROLLER
FREDERICK SEITZ
WILLIAM SHOCKLEY
RICHARD M. SUTTON
J. H. VAN VLECK
MARK W. ZEMANSKY

Consultant on Printing

MELVIN LOOS

ADMINISTRATIVE STAFF

HENRY A. BARTON, *Director*
WALLACE WATERFALL, *Executive Secretary*
THEODORE VORBURGER, *Advertising Manager*
RUTH F. BRYANS, *Publication Manager*
BETTY H. GOODFRIEND, *Assistant Secretary*
EILEEN B. NEUBERGER, *Circulation Manager*
KATHRYN SETZE *Assistant Treasurer*

The JOURNAL OF APPLIED PHYSICS, published monthly at Prince and Lemon Sts., Lancaster, Pa., is devoted to physics in its role as the science basic to other natural sciences and to the arts and industries. Previous to 1937 this publication was known as PHYSICS. The JOURNAL OF APPLIED PHYSICS publishes editorials and reviews, as well as technical papers of applied physics. Articles appearing in it are indexed regularly, according to their field, by Physics Abstracts, Chemical Abstracts, Engineering Index, Inc., Industrial Arts Index, etc.

Manuscripts should be submitted to Elmer Hutchisson, Editor, Case Institute of Technology, Cleveland 6, Ohio. Unless otherwise stated, submission of a manuscript is a representation that it has been neither copyrighted, published, nor submitted for publication elsewhere. The authors' institutions are requested to pay a **publication charge** of $8 per page which, if honored, entitles them to 100 reprints without further charge. Instructions will be sent with galley proofs.

Subscriptions, renewals, and orders for back numbers should be addressed to the American Institute of Physics, 57 East 55 Street, New York 22, New York.

Subscription Price	U.S. and Canada	Elsewhere
To members of the American Institute of Physics and Affiliated Societies	$10.00	$11.00
To all others	12.00	13.00

Back Number Prices

Yearly back number rate when complete year is available: $14.00. Single copies: $1.25 each.

Changes of address should be sent to the American Institute of Physics; **corrected proofs** should be sent to the Publication Manager, American Institute of Physics. **Advertising rates** supplied on request. Orders, advertising copy, and cuts should be sent to the American Institute of Physics.

Entered as second class matter January 22, 1937, at the Post Office at Lancaster, Pennsylvania, under the Act of March 3, 1879. Accepted for mailing at the special rate of postage provided for in paragraph (d-2), section 34.40, P. L. & R. of 1948, authorized May 2, 1932.

Journal
of
Applied Physics

Volume 24, Number 4 April, 1953

Statistical Criteria for the Detection of Pulsed Carriers in Noise. I

DAVID MIDDLETON*
Cruft Laboratory, Harvard University, Cambridge, Massachusetts
(Received August 5, 1952)

A theory for the detection of pulsed carriers (of constant amplitude) in narrow-band random noise is described, based on several types of optimum tests of a statistical hypothesis against an alternative (i.e., signal and noise *vs* noise). Siegert's concept of the betting curve is introduced, whereby, for a finite integration time, the minimum detectable signal is uniquely defined. Three types of observer are next considered: the Neyman-Pearson, the Ideal, and the Sequential observer, whose properties are determined by the manner in which the test is carried out. For each observer it is verified that the best second detector is a $\log I_0$-rectifier, which in practice is closely approximated by the usual half-wave linear envelope-tracer. In Part II specific betting curves are calculated and the performance of the three observers is analyzed for both the weak (threshold) and strong-signal cases.

1. INTRODUCTION

THE detection of signals in the presence of noise is a central problem of communication theory. Various criteria of system performance have been proposed, among them the signal-to-noise ratio, the deflection criterion (for radar), the amount of second and higher harmonic distortion, etc. In one form or another these all depend on the distribution (with frequency) of the power in the input to and the output from the receiver, and following such typical nonlinear devices as the second detector,† require for their calculation knowledge of the second-order probability density $W_2(X_1, X_2; t)$ of the original noise or signal and noise wave. Here, as before, $W_2 dX_1 dX_2$ represents the *joint* probability that at some initial time ($t_1=0$) the amplitude of the disturbance lies in the interval (X_1, X_1+dX_1) and that at a time t later X falls in the range (X_2, X_2+dX_2). From W_2 one gets by well-known methods[1] the autocorrelation function and the spectral intensity of the interference and signal following detection.

Unfortunately, when these theoretical criteria are related to experiment, one must introduce in appropriate fashion a numerical factor (h'), characteristic of the human observer‡ and the rather loose definition of what constitutes a minimum detectable signal in noise. The precise value of h' cannot then be calculated. This uncertainty in h' lies in the fact that the minimum detectable signal is not uniquely defined: the mathematical description of the observation process is incomplete. A second factor limiting the usefulness of these criteria is that they do not indicate the optimum manner in which the desired information is to be presented. Moreover, measures of system performance like these, which utilize at most the information con-

* This work is a condensation of part of a report prepared by the author as a consultant for the Air Force Cambridge Research Center, and is published by permission, which is hereby gratefully acknowledged.

† Under the usual conditions of operation, where the local oscillation is strong relative to the maximum intensity of the modulated carrier, the converter or first detector in heterodyne systems may be safely considered linear. The incoming signal (and noise) is merely shifted to the new, intermediate frequency, without distortion.

[1] S. O. Rice, Bell System Tech. J. 24, 46 (1945); D. Middleton, Quart. Appl. Math. 5, 445 (1948), Sec. 1. For linear systems the second-order, second moment $\langle X_1 X_2 \rangle$ is sufficient, but in nonlinear cases where the output (I) is related to the input (X) by an expression of the type $I=g(X)$, it is necessary, in order to determine the correlation function $\langle I_1 I_2 \rangle$, to know *all* second-order moments $\langle X_1^k X_2^l \rangle$, or equivalently, the density function $W_2(X_1, X_2; t)$. For a discussion, see also D. Middleton, "Noise and Non Linear Communication Problems," Symposium on Applications of Autocorrelation Analysis to Physical Problems, Woods Hole, Massachusetts, June 13, 14 (1949), U. S. Office of Naval Research (May, 1950).

‡ Strictly speaking, this numerical factor h' is given in terms of a *statistical observer*, who represents the mean performance of a theoretically infinite number of individual observers for the problem at hand. For a detailed account of the role of h' in power and deflection criteria, and the concept of the minimum detectable signal, see J. L. Lawson and G. E. Uhlenbeck, *Threshold Signals*, vol. 24, Massachusetts Institute of Technology, Radiation Laboratory Series (McGraw-Hill Book Company, Inc., New York, 1950), Secs. 7.1 and 7.3.

tained in the second-order probability density for the random process in question, clearly throw away a useful portion of the total available information. For during the "smoothing" or integration period at the observer's disposal, the data is presented either continuously or as a sequence of discrete observations, and thus for best results one must use not only the distribution density governing observations at *pairs* of time-points, separated by intervals $t=t_2-t_1$ seconds in duration, *viz.*, W_2, but the higher order densities $W_3; W_4, \cdots;$ $W_n(X_1; X_2, t_2; \cdots; X_n, t_n)$ as well, which govern the data for triples, quadruples, etc. of time-points. Decisions based on the information contained in these higher order density functions are obviously more reliable than decisions made on the basis of W_2 alone, i.e., on the correlation function and spectrum of the signal and noise following the second detector. An improved theory of the detection process is accordingly needed, which (1), describes quantitatively the method of observation, (2) which defines uniquely the minimum detectable signal and determines the threshold constant h', and (3) which indicates optimum ways of presenting the available information.

Such a theory can be constructed if we observe first that our problem is fundamentally statistical. Accordingly, to determine the presence or absence of a signal in noise is equivalent, in statistical language, to testing the hypothesis (H_0) that noise alone is present against the alternative hypothesis (H_1) of a signal and noise. A decision based on the finite amount of information at our disposal is then subject to two types of error:§ a type I error, of calling noise a signal when really noise alone is present, and a type II error, of considering a signal as noise, when a signal is actually present; the respective probabilities of these errors we denote by α and β. In reaching a decision, the observer now in effect makes a "bet," which in the limit of an infinitely large number of repeated trials has a definite probability of "failure," i.e., a faulty decision. *The process of observation is then completely specified analytically by the way in which the observer uses his data to make his bet.* The probability of a success depends on the ratio a_0 of the input signal strength to a (fixed) amount of interfering noise, and ranges in value from the *a priori* probability of a signal or noise, whichever is the larger, when no signal is actually present, to a maximum value unity for infinitely strong signals, where the bet is always completely "successful."

The relation between the percentage of successes (i.e., correct decisions) and the *input* signal-to-noise ratio a_0 is expressed in terms of a *betting curve*, which is the essential feature of the more fundamental theory of the detection process, for with its help we can now define unambiguously the minimum detectable signal: This is that signal observed in the output of the receiver for which there is some arbitrarily selected percentage of success, 90 percent or 50 percent being values frequently taken in practice. The betting curve then connects at once this observed, minimum detectable signal with the corresponding input signal-to-noise ratio, and in this way are avoided the difficulties of describing precisely what is a signal in noise following rectification and of relating this in turn to the incoming signal and noise. Specifically, if p and q are the *a priori* probabilities of signal and noise and noise alone, i.e., the probabilities that out of an indefinitely large number of trials a fraction p is known to contain a signal and noise and a fraction $q(=1-p)$ is known to contain just noise, the betting curve $W_1(a_0^2; n)$ is clearly given by

$$W_1(a_0^2; n)=p(1-\beta)+q(1-\alpha), \tag{1.1}$$

where n is the sample size and is proportional to the integration time; the probabilities α and β of the type I and II errors are in general functions of a_0^2 and n. The betting curve was first introduced by Lawson and Uhlenbeck,[2] and used by them and their co-workers[3] in the study of A-scope and PPI radar reception. The first calculation of a betting curve was initiated by Siegert[4] in his theory of the Ideal observer, a short and somewhat generalized account of which is included in Part II for comparison with observers using different statistical criteria.

In making a test of a statistical hypothesis in the present case, we require first that the test for a possible signal be carried out at some specified range setting. If this is not the case, and if we should then ask for the *exact* position of a signal response (with 100 percent accuracy), as well as testing for its presence, information would be lost, as Woodward has pointed out,[5] since only a *probability distribution* of position can have meaning in such instances. To ask for the exact position on the basis of necessarily incomplete data is to falsify to a certain extent the existing information. In the present paper, we have essentially eliminated this possibility by fixing the "gate" at which the statistical test is to be made.

Let us now apply these ideas to the specific problem of the detection of a pulsed carrier in normal random noise at some *a priori* selected range setting, a problem which is the principal subject of the present paper. We begin by considering a periodic *cw*-pulse train of fixed maximum intensity, which is received in noise of a certain constant level. The observer, who knows only from *a priori* information that a signal and noise or noise alone is likely to occur at his range gate, is then asked to

§ If an infinite amount of information is available, e.g., the data obtained during an infinite integration period, a signal of *any* intensity can, of course, be detected without error.

[2] J. L. Lawson and G. E. Uhlenbeck, *Threshold Signals* (McGraw-Hill Book Company, Inc., New York, 1950), Chapters 7, 8, Secs. 8.3, 8.6, 8.8.

[3] Ashby, Josephson, and Sydoriak, *Signal Threshold Studies*, Naval Research Lab. Report R-3007, December 1, 1946, Appendix I.

[4] A. J. F. Siegert, see preface of reference 2.

[5] P. M. Woodward, Telecommunications Research Establishment (England) Journal, (January, 1951), p. 32. See also the concluding paragraphs of this paper (Part II).

decide on the basis of n observations (equally spaced in time by an amount equal to the pulse repetition period) of the *envelope* of the narrow-band noise and pulsed carrier, whether or not a signal is actually present at the chosen range setting. A *one-point theory*[6] is assumed, which does not take into account the actual shape of the filtered pulses. Only the deflection at the maximum of the envelope, after IF and video filtering,‖ is considered, and the observer accordingly chooses his observation epoch, t, to coincide with this time point. We remark also that only the one-position, or "off-on" problem is considered here, in which the observer's attention is confined to a single range point or (narrow) gating interval. The present approach, however, may be extended to multipositional cases, as Siegert has done in his theory of the Ideal observer,[2,4] and as Woodward has indicated in a recent memorandum.[5]

Our problem is now easily stated in statistical terms: we observe as before that the two types of error I, II are possible. We next must test the statistical hypothesis H_0 of noise alone against the alternative hypothesis H_1 that noise *and* a signal are present. An optimized¶ decision can then be reached by any one of the following three procedures, which describe analytically the observation process:

(1) Fix the probability α and the number of observations n (fixed integration time), and then *minimize* the probability β of a type II error. This is the Neyman-Pearson test of the statistical hypothesis H_0 against H_1 for fixed probability of a type I error, a test which has been shown to be the best possible for a single alternative under these conditions.[7] We shall accordingly call an observer operating in this way a *Neyman-Pearson observer*.**

(2) Divide the observation space for the n data elements in such a way that for fixed n the probability of a correct decision is a maximum, for all possible signal strengths. Here the *sum* $(\alpha+\beta)$ is minimized, and consequently the betting curve, Eq. (1.1) becomes a maximum for a given value of the input signal-to-noise ratio a_0, to satisfy this requirement. An observer who does this is called an *Ideal observer*, and as such was introduced originally by Siegert.[8]

(3) Fix both the probabilities of type I and II errors, leaving the integration time (n) adjustable. Continue the observations until certain limits, depending on α and β, are exceeded for the first time, at which point accept or reject the hypothesis H_0 on the basis of this sequential test. An observer who follows this procedure we shall call a *Sequential observer*.

The Neyman-Pearson criterion has recently been applied to this problem by a number of investigators,[9–12] but until now the essential features of the betting curve and the calculation of the minimum detectable signal have not been specifically introduced. The present preliminary discussion of the theory of the Sequential observer is also believed to be new, but it should be remembered that Siegert's original treatment of the Ideal observer forms the starting point for the present discussion.

2. STATISTICAL PRELIMINARIES

Before a theory of these optimum observers (1–3) can be constructed, it is necessary to consider briefly a few of the underlying statistical ideas. We begin with the Neyman-Pearson test.

(a) Neyman-Pearson Test of an Hypothesis against a Single Alternative[7,13]

Suppose it is known that a random variable R has an nth order probability density $W_n(R_1; R_2, t_2; \cdots R_n, t_n | \lambda)$, where λ is a parameter associated with the distribution. Under one set of conditions λ has the value λ_0 and under all other circumstances λ takes the value λ_1, such that

[6] Reference 2, Sec. 7.3.

‖ An essentially infinite, uniform video response is assumed here; otherwise, one has to treat the far more difficult problem of determining these distributions of the envelope after second detection *and* a finite video filter. We note also that more refined theories are possible, which take progressively greater account of pulse shape and which use in turn two, three, or m time-points per observation, but these higher order descriptions do not appear necessary to explain the experimental observations to date (reference 6) and are therefore not examined here. Furthermore, the loss of information occasioned by the 1-point theory is ignorable from a practical standpoint, provided the gate is equal to or less than the pulse width.

¶ Although each decision is the optimum one for its own indicated "observation" process, as we shall see presently (in Part II, Sec. 7), it is not true that they are equally good. This is to be decided by a suitable comparison, the precise nature of which is described in Part II, Sec. 7.

[7] J. Neyman and E. S. Pearson, Trans. Roy. Soc. (London) **A231**, 289 (1931); J. Neyman, J. Roy. Stat. Soc. **105**, 292 (1942).

** An alternative procedure is to fix β and minimize α instead, for a constant integration time. But it is evident by the same argument (Sec. 2) that this is essentially equivalent to (1) above, where now $\beta_{\min}$ becomes β, and α is $\alpha_{\min}$ (their numerical values being unchanged by the alteration in viewpoint). However, since the parameter involved in H_1 (i.e., the input signal-to-noise ratio a_0) is usually not known, it is not possible to specify this fixed β $(=\beta(a_0))$ and so α is chosen instead, because it is derived from a distribution independent of this parameter.

[8] Reference 2, Sec. 7.5.

[9] H. Hanse, "The Optimization and Analysis of Systems for the Detection of Pulsed Signals in Random Noise," Doctoral dissertation, Massachusetts Institute of Technology, January, 1951. Some experimental data is also given here.

[10] M. Schwartz, "A Statistical Approach to the Automatic Search Problem," Doctoral dissertation, Harvard University, June, 1951.

[11] J. Marcum, Rand Corporation Reports, (C)-RA-15061, December 1, 1947; (R)-R-113, July 1, 1948, Mathematical Appendix.

[12] Several interesting papers (by D. L. Drukey, Hughes Aircraft, and by T. G. Slattery) incorporating the Neyman-Pearson ideas of testing statistical hypotheses and applying them to the problem of the pulsed carrier in noise have subsequently been presented at the Institute of Radio Engineers National Convention, March, 1952. See also, T. G. Slattery, Proc. Inst. Radio Engrs. **40**, 1232 (1952). However, a discussion of the Ideal and Sequential observers, the notion of the betting curve, and the precise calculation of the minimum detectable signal have not been considered.

[13] For a detailed discussion, see for example A. M. Mood, *Introduction to the Theory of Statistics* (McGraw-Hill Book Company, Inc., New York, 1950), Chaps. 12, 15; also, H. Cramér, *Mathematical Methods of Statistics* (Princeton University Press, Princeton, 1946), Chaps. 30–37. A more elementary treatment is given in Chap. 11 of P. G. Hoel's, *Introduction to Mathematical Statistics* (John Wiley and Sons, Inc., New York, 1947).

$W_n(R_1; R_2, t_2; \cdots R_n, t_n|\lambda_0)$ and $W_n(R_1; R_2, t_2; \cdots; R_n, t_n|\lambda_1)$ are different functions. Thus in the present instance λ_1 is a_0, where $a_0 = A_0/(2\psi)^{\frac{1}{2}}$ and A_0 is the (peak) carrier amplitude and ψ the mean-square amplitude of the noise, both measured at the same point somewhere before the second detector. Furthermore, λ_0 is zero, for the case where noise alone is the input to the second detector, and $R_1, R_2, \cdots R_n (\geq 0)$ represent n (successive) values of the *envelope* of the incoming wave, observed at the time points $0, t_1, t_2 \cdots t_n$, respectively.

We now wish to test on the basis of any one set of n observations of the envelope $R_1', \cdots, R_n'$ the hypothesis H_0 that λ has the value $\lambda_0 = 0$ against the alternative H_1 that λ is equal to $\lambda_1 = a_0 (> 0)$. Here H_0 is frequently called the *null hypothesis*, which is usually associated with those conditions which are known to occur most of the time,†† and whose validity for the n data elements in question is to be tested. Clearly, rejection of H_0 automatically implies acceptance of H_1 and vice versa, since there is but one alternative.

Our test for accepting or rejecting H_0 requires that we divide the observation space into two parts, one of which is associated with H_0 and the other with the alternative H_1. There are then two possible kinds of error:

(1) *Errors of the first kind, or type I errors*—H_0 is rejected when it is in fact true. The probability of a type I error is given by

$$\alpha = \int \cdots \int_{\Gamma''} W_n(R_1; R_2, t_2; \cdots; R_n, t_n|\lambda_0) \times dR_1 \cdots dR_n, \quad (2.1)$$

where Γ'' is the region of the n-fold observation space $R_1, \cdots, R_n$, which is associated with the alternative hypothesis H_1, and which is specified by the manner in which the test is to be made.

(2) *Errors of the second kind, or type II errors*—here H_0 is accepted, although H_1 is really true. The probability of a type II error is then

$$\beta = \int \cdots \int_{\Gamma'} W_n(R_1; R_2, t_2; \cdots; R_n, t_n|\lambda_1) \times dR_1 \cdots dR_n, \quad (2.2)$$

where $\Gamma = \Gamma' + \Gamma''$ includes all of the n-dimensional sample space for $R_1, \cdots, R_n$. Here Γ' and Γ'' are, respectively, called the *acceptance* and the *critical* regions, and $1-\beta$ is known as the *power* of the test, while α is often called the *significance level*. We observe now that the construction of a test is determined by the manner in which the observation space $\Gamma = \Gamma' + \Gamma''$ is divided into the two regions Γ' and Γ''. H_0 is accepted or rejected on the basis of whether or not our n-fold data point $(R_1', \cdots, R_n')$ lies in Γ' or Γ''. Although this division is quite arbitrary, some tests are better than others, and it is therefore reasonable to ask how the division should be made so that one obtains the best test. Usually for small probabilities of type I errors the probabilities of type II are large, and vice versa. For this reason and the fact that the distribution from which the probability β is calculated is not usually specifically known, Neyman and Pearson fix α arbitrarily, and then adjust the critical region Γ'' to minimize the probability β, i.e., to make Γ' a *most powerful region*.

A simple application of the calculus of variations now shows that under these conditions the acceptance and critical regions are separated by the surface

$$\frac{pW_n(R_1''; R_2'', t_2; \cdots; R_n'', t_n|\lambda_1 = a_0)}{qW_n(R_1''; R_2'', t_2; \cdots; R_n'', t_n|\lambda_0 = 0)} = K = K(\alpha; R_1'', \cdots R_n''; a_0), \quad (2.3)$$

where $R_l''(l = 1, \cdots n)$ are n values of the random variable R so chosen that Eq. (2.3) is true, once α has been given. The variational operation is easily carried out in a formal way: When p and q are different, we observe that we have to extremize $p\beta = p \int_{\Gamma'} W_n(\Gamma|\lambda_1)\mathbf{d\Gamma}$ with respect to a fixed $q\alpha = q \int_{\Gamma''} W_n(\Gamma|\lambda_0)\mathbf{d\Gamma}$, where by way of notation $\mathbf{d\Gamma} = dR_1 dR_2 \cdots dR_n$, $\Gamma = (R_1, R_2 \cdots, R_n)$, and the form of W_n is at least given. Our problem now is to divide the region $\Gamma = \Gamma' + \Gamma''$ in such a way that β is minimized. We can write $p\beta + \Lambda_n \alpha q$ equal to some constant, where Λ_n for the moment is an undetermined multiplier, so that

$$\delta(p\beta + \Lambda_n \alpha q) = 0, \quad (2.4a)$$

or

$$\delta \int_{\Gamma_1}^{\Gamma_2} pW_n(\Gamma|\lambda_1)\mathbf{d\Gamma} + \Lambda_n \delta \int_{\Gamma_2}^{\Gamma_3} qW_n(\Gamma|\lambda_0)\mathbf{d\Gamma} = 0. \quad (2.4b)$$

Here $\Gamma_2 = (R_1'', R_2'', \cdots, R_n'')$ represents the coordinates of the surface dividing the regions Γ' and Γ'', and Γ_1 and Γ_3 are, respectively, the lower and upper bounds to the whole observation space: $\Gamma_1 = (0, 0, \cdots, 0)$, and $\Gamma_3 = (\infty, \cdots, \infty)$. Carrying out the indicated variations of the boundary Γ_2, we have

$$\int_{\Gamma_2}^{\Gamma_2 + \delta\Gamma} pW_n(\Gamma|\lambda_1)d\Gamma + \Lambda_n \int_{\Gamma_2 + \delta\Gamma}^{\Gamma_2} qW_n(\Gamma|\lambda_0)d\Gamma = 0, \quad (2.5a)$$

or

$$\delta\Gamma\{pW_n(\Gamma|\lambda_1) - \Lambda_n qW_n(\Gamma|\lambda_0)\} = 0, \quad (2.5b)$$

with the result that, since $\delta\Gamma$ is arbitrary,

$$\Lambda_n'' = \frac{pW_n(R_1'' \cdots R_n''|\lambda_1)}{qW_n(R_1'' \cdots R_n''|\lambda_0)} = K. \quad (2.6)$$

†† Since the detection of a signal in noise is basically the threshold problem of distinguishing a *weak* signal against the noise background, it is natural to choose $H_0(\lambda = \lambda_0 = 0)$, rather than $H_1(\lambda = \lambda_1 = a_0)$ as the null hypothesis.

This is the desired condition, Eq. (2.3). Then β is minimized, once $\Gamma''=(R_1''\cdots R_n'')$, for our choice of α, has been selected. [Clearly, one gets the same result Eq. (2.5); Eq. (2.6) if β is fixed and α is minimized.] Here Λ_n is called the *likelihood ratio*. Consequently,

$$\Lambda_n=\frac{pW_n(R_1;R_2,t_2;\cdots;R_n,t_n|\lambda_1)}{qW_n(R_1;R_2,t_2;\cdots;R_n,t_n|\lambda_0)}. \tag{2.7}$$

We see, therefore, for a set of observations $R_1=R_1'$, $R_2=R_2'$, $\cdots$, $R_n=R_n'$, such that the likelihood ratio above exceeds K, that H_0 is then rejected and H_1 accepted, while if this ratio is less than K the reverse is the case. Decisions made on the basis of Eq. (2.3) are the best decisions that one can make, and this is the best test, in the Neyman-Pearson sense. Figure 1 illustrates some of these remarks.

Although we have tacitly assumed that $\lambda_1=a_0$ has only one particular value for all tests of n observations, the optimum character of the present test is sometimes not altered if other choices of a_0 are made. This is a situation, however, which is not generally true of every statistical test, but only of those for which the critical region Γ'' remains unchanged when $\lambda=\lambda_1=a_0$ assumes all possible values. This in turn requires that the likelihood ratio Λ_n can be expressed in the form $G(a_0)H(R_1\cdots R_n)$ for all a_0. As we can see from Eq. (2.16) following, such a factorization of Λ_n is not realized for the type of noise examined here, although it can be approximated if a_0^2 is sufficiently small. If Λ_n can be factored in this way, however, the Neyman-Pearson criterion is, therefore, a *uniformly most powerful* test, no matter what other tests (for a given α) are tried.[14] This has the further important consequence, as Neyman and Pearson have shown,[15] that a uniformly most powerful test also remains the optimum one, no matter what the *a priori* probabilities of signal and of noise may be.

(b) Sequential Test of an Hypothesis against a Single Alternative

In the Neyman-Pearson theory outlined above the best test for accepting or rejecting $H_0(\lambda=\lambda_0=0)$ *versus* $H_1(\lambda=\lambda_1=a_0)$ is based on minimizing the probability β of the type II error for a fixed probability of a type I error and fixed sample size. One can instead set initially both the probabilities of the type I and type II errors. Now, however, sample size cannot be chosen in advance, but is determined in the following way: For a best test, the Neyman-Pearson theory is again invoked, but with the sample size n for the moment unspecified. If next we plot the maximum power $(1-\beta_{\min})$ of the test for each corresponding value of n and the given probability of a type I error α, we obtain a (broken) curve, which increases monotonically with increasing size, since the larger the sample the more precisely can the correct decision of accepting H_0 or H_1 be made. Then for the α and β (chosen) here, one has only to read off from such a curve the now specified sample size n for this optimum test. Note also that in each case the critical region Γ'' is a function of n, changing when the sample size is altered.

Let us now suppose that for a given α and β, and the associated probability densities $W_n(R_1;R_2,t_2;\cdots R_n,t_n|\lambda_1)$, $W_n(R_1;R_2,t_2;\cdots;R_n,t_n|\lambda_0)$, the above process is followed and n is found to be, say, 90. In making these 90 observations to test H_0 against H_1 it is possible that among the first few observations there may be one or more so far into the acceptance region Γ' that eventual rejection of H_0 is out of the question. Further observations are, therefore, a waste of time. It is even more probable that the information from the first 20, 30, or 40 observations is sufficient to allow us to accept or reject H_0, subject to the predetermined values of α and β. Thus, by acting on the information *as it is obtained sequentially*, one can make savings in sample size, with the same controls used in the Neyman-Pearson test. This saving may be considerable, sometimes as much as 50–70 percent over the fixed sample procedure, depending of course on α and β. Clearly, it may be profitable, where limited amounts of time are available to the observers, to design a sequential test[16] of H_0 against H_1. At this point we remark that such sequential tests are never *less* effective than the corresponding Neyman-Pearson procedures, i.e., they never require a larger sample for the same value of the probabilities α and $\beta(=\beta_{\min\text{-NP}})$, since the Neyman-Pearson test (of sample size n) can always be regarded as a special case of sequential sampling in which the test terminates after exactly $n=\langle n\rangle$ samplings.

The design and operation of such a sequential test is quite straightforward. We start with the likelihood ratio

$$\Lambda_{n(s)}=\frac{pW_{n(s)}(R_1;R_2,t_2;\cdots;R_{n(s)},t_{n(s)}|\lambda_1)}{qW_{n(s)}(R_1;R_2,t_2;\cdots;R_{n(s)},t_{n(s)}|\lambda_0)} \tag{2.8}$$

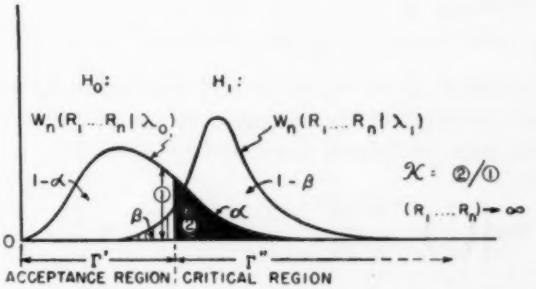


FIG. 1. A schematic representation of the probability densities in a Neyman-Pearson test of H_0 *vs* H_1.

[14] H. Cramér, *Mathematical Methods of Statistics* (Princeton University Press, Princeton, 1946), Chap. 35, and Bayes' theorem, Chap. 34, p. 508.

[15] J. Neyman and E. S. Pearson, Proc. Cambridge Phil. Soc. **29**, 492 (1933).

[16] The methods of sequential analysis are due primarily to Wald, *Sequential Analysis* (John Wiley and Sons, Inc., New York, 1947). For some of the principal results, see also Chap. 15 of Mood, reference 13.

and choose two positive numbers A_1 and A_2, depending on α and β, such that $A_2>1$ and $0<A_1<1$. As each observation in the course of a run is made, the ratios $\Lambda_1, \Lambda_2, \Lambda_3, \cdots$ are formed, as long as

$$A_1(\alpha;\beta)<\Lambda_{n(s)}<A_2(\alpha;\beta). \tag{2.9}$$

As soon as $\Lambda_{n(s)}$ either is equal to or exceeds A_2, or is equal to or less than A_1, we respectively reject H_0 in favor of H_1, or we accept it, terminating the test. It is not too difficult to show[13] that in the case of independent samples the test always terminates for some finite $n(s)$; this is also true in many physical cases even where there is correlation between successive observations. Then, by an argument discussed by Wald,[16] and Mood[13] (Chap. 15), the upper and lower bounds A_2 and A_1 can be expressed in terms of the probabilities of the type I and II errors according to

$$A_1=\beta/(1-\alpha);\quad A_2=(1-\beta)/\alpha. \tag{2.10}$$

This is precise for *continuous* sampling, where $n(s)$ is regarded as a continuous function, but the error introduced when Eq. (2.10) is used in discrete sampling is small, provided α and β are both less than $\frac{1}{2}$. Here α and β are given by Eqs. (2.1) and (2.2), as before, if n is replaced by $n(s)$. Our test of H_0 *vs* H_1 for fixed α and β accordingly consists of making $n(s)$ observations $R_1=R_1'$, $R_2=R_2'$, $\cdots$, $R_{n(s)}=R_{n(s)}'$ until the likelihood ratio, Eq. (2.8) falls for the first time outside the limits indicated by Eq. (2.9). Equation (2.9) then determines $n(s)$, which is of course a random variable when considered over the ensemble of all possible runs.

A satisfactory comparison of the sequential test with tests of fixed sample size requires that we know, at least, the mean sample size $\langle n(s)\rangle=\langle n\rangle$ and variance‡‡ $\langle n(s)^2\rangle-\langle n(s)\rangle^2=\sigma_s^2$. So far, these have been obtained only for independent observations, which, however, is the usual case of our present problem, since the noise is here assumed uncorrelated from pulse to pulse, Sec. 3. We have specifically[16] (Appendix I, Air Force Cambridge Research Center Report No. E-5091)

$$\langle n\rangle=\langle Z_{n(s)}\rangle/\langle z\rangle;\quad \langle n^2\rangle-\langle n\rangle^2=\{\langle Z_{n(s)}\rangle(\langle z^2\rangle-\langle z\rangle^2)/\langle z\rangle-[\langle Z_{n(s)}{}^2\rangle-2\langle z\rangle\langle nZ_{n(s)}\rangle+\langle Z_{n(s)}{}^2\rangle]\}/\langle z\rangle^2, \tag{2.11}$$

Here $W_1(R|\lambda)$, $\lambda=\lambda_1$ or $\lambda=\lambda_0$, is the first-order probability density for the sample envelope of the narrow-band noise and signal wave, and z is given by

$$z=\log\left\{\left(\frac{p}{q}\right)^{1/n}\frac{W_1(R|\lambda_1)}{W_1(R|\lambda_0)}\right\},\quad\text{and}\quad Z_{n(s)}=\sum_{k=1}^{n(s)}z_k, \tag{2.12}$$

in which $n(s)$ is the smallest integer for which $Z_{n(s)}$ does *not* satisfy the fundamental inequalities Eq. (2.9) of the test, which become now

$$\log A_1(\alpha,\beta)<Z_{n(s)}=\sum_{k=1}^{n(s)}\log\left\{\left(\frac{p}{q}\right)^{1/n}\frac{W_1(R_k|\lambda_1)}{W_1(R_k|\lambda_0)}\right\}<\log A_2(\alpha,\beta). \tag{2.13}$$

The probability density $w(z)$ governing the z_k is derived from Eq. (2.12) with the help of the *true* distribution density for R, which for threshold signals is taken to be $W_1(R|\lambda_0=0)$. For strong signals, on the other hand, one takes as the true distribution density $W_1(R|\lambda_1=a_0)$.

The precise calculation of $Z_{n(s)}$ depends on $W_s[n(s)]$, which is not generally known. However, a simple approximation can easily be obtained, as discussed by Mood.[16] In the threshold case, where $a_0=0$, we have approximately for the distribution density $W'[Z_{n(s)}]$

$$W'(Z_{n(s)})\doteq\alpha\delta[Z_{n(s)}-\log A_2]+(1-\alpha)\delta[Z_{n(s)}-\log A_1],\quad(\text{threshold: }\lambda=\lambda_0), \tag{2.14a}$$

where δ is the well-known Dirac delta-function. For the case of strong signals, we obtain correspondingly

$$W'(Z_{n(s)})=(1-\beta)\delta(Z_{n(s)}-\log A_2)+\beta\delta(Z_{n(s)}-\log A_1)\quad(\text{strong signal: }\lambda=\lambda_1), \tag{2.14b}$$

and so from Eq. (2.11) we get at once in these two situations

$$\langle n(s)\rangle=(\alpha\log[(1-\beta)/\alpha]+(1-\alpha)\log[\beta/(1-\alpha)])/\langle z\rangle;\quad(\lambda=\lambda_0) \tag{2.15a}$$

$$=((1-\beta)\log[(1-\beta)/\alpha]+\beta\log[\beta/(1-\alpha)])/\langle z\rangle;\quad(\lambda=\lambda_1), \tag{2.15b}$$

respectively. Specifically, the probability density $W_1(R|\lambda)$ for the envelope R of noise alone or of a carrier and noise is given by[17]

$$W_1(R|\lambda_0=0)=(R/\psi)e^{-R^2/2\psi},\ R>0;\ =0,\ R<0, \tag{2.16a}$$

$$W_1(R|\lambda_1=a_0)=(R/\psi)e^{-a_0{}^2-R^2/2\psi}I_0(a_0R[2/\psi]^{\frac{1}{2}}),\quad R>0;\ =0,\ R<0;\ r_0=R_0/[2/\psi]^{\frac{1}{2}}, \tag{2.16b}$$

where ψ is the $r-m-s$ input noise amplitude to the second detector, A_0 is the peak carrier amplitude (when the carrier is on), measured at the same point in the receiver as the noise, and I_0 is a modified Bessel function of the first kind, zeroth order. For the threshold cases ($a_0{}^2<1$) and equal *a priori* probabilities $p=q$, when noise described by Eq. (2.16) is present, we find that§§

$$\langle z\rangle\doteq-a_0{}^4/2;\quad\langle z^2\rangle\doteq a_0{}^4, \tag{2.17a}$$

‡‡The variance is needed to give an estimate of the spread of the distribution density $W_s(n(s))$, which, unfortunately, does not appear obtainable in the general case.

[17] S. O. Rice, Bell System Tech. J. 24, 46 (1945), Part III; D. Middleton, Quart. Appl. Math. 5, 445 (1948), for example.

§§ Full details, including the cases $p\neq q$ are described in Appendix I of Air Force Cambridge Research Center Report No. E-5091.

while in the strong-signal case we obtain

$$\langle z\rangle \simeq +a_0^2; \quad \langle z^2\rangle \simeq a_0^4. \tag{2.17b}$$

In a similar way $\langle n(s)^2\rangle - \langle n(s)\rangle^2$ can be determined. [The details of the derivation of $\langle n\rangle$ and $\langle n^2\rangle - \langle n\rangle^2$, Eq. (2.11), are given in Appendix I, Air Force Cambridge Research Center Report No. E-5091.] These results aid us in attempting to compare sequential tests involving independent observations with optimum tests for fixed sample size.

3. THE OPTIMUM SECOND DETECTOR

We now consider the problem in which we are to detect in the video stage of a receiver the envelope of a periodic pulse train in noise. A stationary process is assumed. We note also that when the interfering noise has normal statistics and an original spectral spread wide compared to the RF and IF stages of the receiver, the correlation time¶¶ of this input to the second detector is then small compared to the pulse repetition period T_0. There is then no correlation between the noise parts of the input at successive pulse epochs. For an integration period T containing n pulses ($T=nT_0$), the noise and noise and signal waves at these successive time points belong therefore to a purely random process, which is described completely[18] by the first-order density $W_1(R|\lambda)$, where $\lambda=\lambda_1=a_0$, or $\lambda=\lambda_0=0$, respectively. Accordingly, the general nth order density functions W_n reduce at once to

$$W_n(R_1; R_2, t_2; \cdots; R_n, t_n|\lambda) = \prod_{l=1}^{n} W_1(R_l|\lambda), \tag{3.1}$$

where $W_1(R_l|\lambda)$ is given by Eqs. (2.16a, b).

Our optimum observers base their decisions in appropriate ways on a division of the observation space $R_1\cdots R_l\cdots R_n$. The likelihood ratio K for which this decision is made becomes now for the Neyman-Pearson and Ideal observers, with the help of Eqs. (2.16), (3.1) in (2.7)

$$K=K(R_1''\cdots R''\cdots R_n''; a_0)$$

$$= pe^{-na_0^2}\cdot \prod_{l=1}^{n} I_0(a_0 2r_l'')/q^{1/n}, \quad r_l''=R_l''/(2\psi)^{\frac{1}{2}}. \tag{3.2}$$

The hypothesis of a signal is accepted or rejected if the likelihood ratio for a set $r_l'(l=1\cdots n)$ of observations satisfies the inequalities

$$\log\Lambda_n = \log\mu - na_0^2 + \sum_{l=1}^{n} \log I_0(2a_0 r_l')$$

$$\begin{matrix} > \\ < \end{matrix} \log K \begin{cases} H_1 \text{ accepted: there is a signal} \\ H_1 \text{ rejected: there is only noise,} \end{cases} \tag{3.3a}$$

where $\mu=p/q$, or equivalently, if

$$\sum_{l=1}^{n} \log I_0(2a_0 r_l') \gtrless K'(=\log K + na_0^2 - \log\mu). \tag{3.3b}$$

In the case of the Sequential observer, it is clear from Eqs. (2.9) and (2.13) that the observation space is divided into three regions by A_1 and A_2; here A_1 and A_2 take the role of K above, and one can write specifically as the condition of the test in this instance

$$K_1' \equiv -\log\mu + n(s)a_0^2 + \log A_1(\alpha, \beta)$$

$$< \sum_{l=1}^{n(s)} \log I_0(2a_0 r_l')$$

$$< -\log\mu + n(s)a_0^2 + \log A_2(\alpha, \beta) \equiv K_2'. \tag{3.4}$$

The important feature of Eq. (3.4) is that the $\log I_0$ structure of the sample remains unchanged (regardless of the *a priori* probabilities p and q).

The results, Eqs. (3.3) and (3.4), immediately answer the question of the best way for our optimum observers to extract the envelope from the IF wave, if post-detector integration is used (as it must be here, since coherence of the carrier from pulse to pulse is not ordinarily possible): *We see that the optimum detector is a half-wave nonlinear device with the dynamic response,*

$$I(t)_{\text{out}} = C\log I_0(A_0 R(t)/\psi), \quad R>0, \quad =0, \quad R<0, \tag{3.5}$$

as shown in Fig. 2. For n pulses we simply add the outputs and accept or reject the hypothesis of a signal according to Eqs. (3.3) and (3.4), for the appropriate K, K_1', K_2'. The optimum character of the $\log I_0$-detector was first pointed out in a published article by Woodward and Davies,[19] and independently somewhat later by Hanse[9] and by Schwartz.[10] When the signal is weak relative to the noise, one has essentially a half-

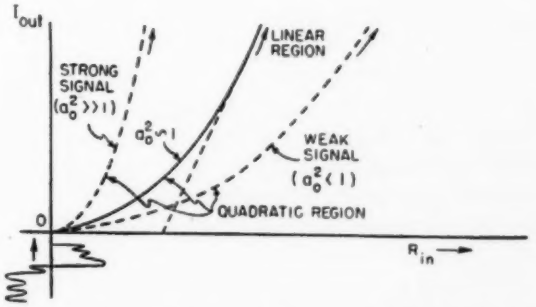


FIG. 2. Optimum (dynamic) characteristics of the second detector for the demodulation of pulsed carriers in normal random noise.

¶¶ This is the time required for the correlation function of the noise, here determined essentially by the over-all filter characteristics of the IF stage, to fall to one-half its maximum value at $t=t_2-t_1=0$.

[18] M. Wang and G. E. Uhlenbeck, Revs. Modern Phys. **17**, 323 (1945); also, D. Middleton, Quart. Appl. Math. **5**, 445 (1948).

[19] P. M. Woodward and I. L. Davies, *A Theory of Radar Information*, Phil. Mag. **41**, 1001 (1950). See also P. M. Woodward, Proc. Inst. Radio Engrs. **39**, 1521 (1951). (Actually, Markum [reference 11], appears to have shown this first in an unpublished report, Rand Corporation [(C)-RA-15061 and (R)-R-113].)

wave *quadratic* detector, namely

$$I_{out}=C\log I_0(A_0R_{in}/\psi)\doteq CA_0^2R_{in}^2/4\psi^2=Ca_0^2r_{in}^2, \quad (a_0^2r_{in}^2\ll 1), \quad (3.6a)$$

while for strong signals, the optimum detector behaves like a half-wave *linear* rectifier (with ignorable bias $\sim\log 2a_0r$), i.e.,

$$I_{out}=C\log I_0(A_0R_{in}/\psi)\simeq 2Ca_0r_{in}, \quad (a^2r^2\gg 1), \quad (3.6b)$$

see Fig. 2. The phenomenon of *modulation suppression*[20,21] (whereby a weak input signal in the presence of noise is made even weaker in the output following a nonlinear device) insures that for threshold reception a linear, half-wave detector behaves almost exactly like a half-wave quadratic device. This allows us to use the simple quadratic approximation, Eq. (3.6a), for the detector's law, in the important threshold cases. When the signal is strong, on the other hand, the optimum $\log I_0$-detector reduces effectively to the actual half-wave linear detector used in practice, Eq. (3.6b), showing that this actual linear detector is very close to the optimum for these observers, no matter what the relative intensities of signal and noise. The extensive mathematical difficulties of calculating betting curves with the $\log I_0$-detector can, therefore, be safely avoided if one assumes throughout the analysis a half-wave linear characteristic. An experimental confirmation of this has been obtained by Hanse[9] for the Neyman-Pearson observer.

The optimum character of the $\log I_0$-device is perhaps not too surprising when we recall that in each instance decisions are based on the likelihood ratio Λ_n, which, in turn, depends on the specific statistical properties of the noise and signal, and these are here the same for all three observers. In fact, given these statistics, the optimum detector is always going to take the $\log I_0$ form for any optimum system of distinguishing between noise alone and a signal and noise, since the likelihood ratio contains all the *a priori* knowledge available about the signal and noise, regardless of how this information is to be subsequently processed. What is perhaps more remarkable is that the method of integration used for so long, namely *adding* the successive pulses following a linear envelope detector, is so close to the optimum indicated for these analytical observers. Note also that although the *gain* of the ideal detector depends explicitly on signal level ($\sim a_0$), the argument of the preceding paragraph makes it clear that the usual half-wave linear detector of *fixed* gain (independent of a_0) is sufficiently close for an optimum performance.

Part II follows, in which the ideas described in Part I are applied to the specific calculations of betting curves and the determination of the minimum detectable signal for the three observers discussed here.

[20] J. H. Van Vleck and D. Middleton, J. Appl. Phys **17**, 940 (1946)
[21] D. Middleton, Proc. Inst. Radio Engrs. **36**, 1467 (1948).

Statistical Criteria for the Detection of Pulsed Carriers in Noise. II*

DAVID MIDDLETON
Cruft Laboratory, Harvard University, Cambridge, Massachusetts
(Received August 5, 1952)

In Part I a statistical theory for the detection of a pulsed carrier in noise has been described, where specifically three types of observer are defined: (1) the Neyman-Pearson, (2) the Ideal, and (3) the Sequential observer. In Part II specific betting curves are determined for the weak-signal cases, and some approximate curves for strong signals are also included. The input signal-to-noise (power) ratio $(a_0^2)_{\min}$ corresponding to the minimum detectable signal at the output (with an arbitrary probability of a correct decision) is found, as expected, to vary as $n^{-\frac{1}{2}}$ for weak signals and as n^{-1} for strong, where n=number of pulses integrated. Furthermore, comparison of observers on the basis of the minimum detectable signal reveals the Neyman-Pearson to be better than the Ideal, by an amount 0 (0.5 db). On the basis of the assumed approximate distribution $w_s(n_s)$ of sample size, the Neyman-Pearson also appears superior to the Sequential observer, but as indicated in Part I, the latter must always be as good as, if not better than the former. The discrepancy is the result of the approximate nature of the assumed distributions. A lower bound on the errors thus introduced in the calculations of the $(a_0^2)_{\min}$ is accordingly provided by the extent to which the Sequential appears *worse* than the Neyman-Pearson observer for equal probabilities of type I and II errors α and β. For $w_s(n_s)=\delta(n_s-\langle n\rangle)$ this error is reasonably small 0 (1.5 db), but for a Gaussian distribution of n_s the effect is 0 (3.0 db) at least. Included also are an analysis of performance and some of the factors affecting it. The paper terminates with a brief account of a statistical theory of the observation process and its relation to the various observers considered here.

INTRODUCTION

IN the preceding three sections (Part I) we have outlined a theory for the detection of a pulsed carrier wave in normal random noise.[1] The analysis provides in a unique way a definition of the minimum detectable signal and takes into account all the available *a priori* and *a posteriori* information, unlike most of the earlier treatments, which have been based essentially on considerations of input and output power and spectra alone, and which leave open the practical question of what precisely constitutes a minimum detectable signal.[2] One of the chief features of the present theory is the concept of the betting curve,[3] defined by the method of observation, and from which in turn the minimum detectable signal is obtained.

The purpose of the present paper (Part II) is first to determine explicitly the betting curves for the Neyman-Pearson, Ideal, and Sequential observers described in Part I, and then to compare the performance of these observers. We begin by considering the Neyman-Pearson observer.

4. THE NEYMAN-PEARSON OBSERVER

Here we wish to determine the betting curve for the observer who holds his probability α of a type I error fixed and minimizes his probability β of a type II error for a constant integration period containing n pulses. We begin by considering a large number of independent trials, during each of which the signal (of given intensity) is observed against a noise background of constant intensity. This signal is presented to the observer, as the number of trials becomes indefinitely great, a fraction p of the time on the average, while for a fraction $q=1-p$, only noise is present at the single range setting of our "off-on" experiment. The Neyman-Pearson observer then makes n observations (following a half-wave *linear* second detector, *vide* Sec. 3, Part I) in the course of each trial, generating an ensemble of trials each of which is nT_0 seconds in duration. Thus, if p is the *a priori* probability of a signal and q the *a priori* probability of no signal at this range, we can immediately write for the betting curve after n observations

$$W_1(a_0;n)_{NP}$$
$$=p(1-\beta_{\min})+q(1-\alpha) \tag{4.1a}$$
$$=p\int_{\Gamma''\text{ (signal on)}}\cdots\int W_1(R_1|a_0)\cdots W_1(R_n|a_0)dR_1\cdots dR_n$$
$$+q\int_{\Gamma'\text{ (signal off)}}\cdots\int W_1(R_1|0)\cdots W_1(R_n|0)dR_1\cdots dR_n. \tag{4.1b}$$

The observation space Γ for the envelopes $R_1\cdots R_n$ is next divided into two regions Γ', Γ'' when the likelihood ratio Λ_n is equal to K, which as we have seen (Part I, Sec. 2) minimizes the probability β of a type II error for this observer (for a given a_0). Rewriting Eq. (4.1b) as

$$W_1(a_0;n)_{NP}=q\int_0^\infty\cdots\int\{\prod_{l=1}^n W_1(R_l|0)dR_l\}\begin{cases}\left(\frac{p}{q}\right)\prod_{l=1}^n W_1(R_l|a_0)/\prod_{l=1}^n W_1(R_l|0)=\Lambda_n, & \text{in } \Gamma''\\ 1, & \text{in } \Gamma'\end{cases} \tag{4.2}$$

* This paper represents a condensation of a report prepared by the author as a consultant to the Air Force Cambridge Research Center, and is published by permission, which is hereby gratefully acknowledged.

[1] D. Middleton, J. Appl. Phys. 24, 371 (1953).

[2] See, for example, J. H. Van Vleck and D. Middleton, J. Appl. Phys. 17, 940 (1946), and the discussion in Chapters 7 and 8 of *Threshold Signals* (McGraw-Hill Book Company, Inc. New York, 1950) by J. L. Lawson and G. E. Uhlenbeck.

[3] J. L. Lawson and G. E. Uhlenbeck, *Threshold Signals* (McGraw-Hill Book Company, Inc., New York, 1950). See, in particular, section 7.5 and the preface.

we observe then that if $\Lambda_n > K$, H_1 is accepted—there is a signal; while if $\Lambda_n < K$, H_0 is accepted instead—only noise is present.

It is convenient now to introduce the functions

$$P_n(x; a_0) = \int_0^\infty \cdots \int \{\prod_{l=1}^n W_1(R_l|a_0) dR_l\} \delta(x - \log\Lambda_n), \quad (4.3a)$$

$$Q_n(x; a_0) = \int_0^\infty \cdots \int \{\prod_{l=1}^n W_1(R_l|0) dR_l\} \delta(x - \log\Lambda_n), \quad (4.3b)$$

where for the second detector and the normal statistics assumed here

$$\log\Lambda_n = \log\mu - na_0^2 + \sum_{l=1}^n \log I_0(R_l a_0[2/\psi]^{\frac{1}{2}}); \qquad \log\mu \equiv \log(p/q). \quad (4.4)$$

The betting curve, Eq. (4.1), can now be written

$$W_1(a_0; n)_{NP} = p\int_{\log K}^\infty P_n(x; a_0)dx + q\int_{-\infty}^{\log K} Q_n(x; a_0)dx. \quad (4.5)$$

The fixed probability of a type I error and the minimized probability of a type II error are therefore given by

$$\alpha = 1 - \int_{-\infty}^{\log K(\alpha;\ a_0;\ n)} Q_n(x; a_0)dx = \text{constant (independent of } a_0), \quad (4.6a)$$

$$\beta_{\min} = 1 - \int_{\log K}^\infty P_n(x; a_0)dx. \quad (4.6b)$$

The critical value K of the likelihood ratio for which the observation space is divided into the "off"- and "on"-regions is determined from Eq. (4.6a). This value is then substituted into Eq. (4.6b) to give us $\beta_{\min}$, and finally the desired betting curve, Eq. (4.5).

Before calculating $W_1(a_1; n)_{NP}$ explicitly, let us examine some of the properties of P_n and Q_n. It is at once evident from Eqs. (4.3a, b) that

$$\int_{-\infty}^\infty P_n(x; a_0)dx = \int_{-\infty}^\infty Q_n(x; a_0)dx = 1, \text{ and } P_n(x; 0) = Q_n(x; 0). \quad (4.7)$$

In the extreme case of a vanishingly small signal, we obtain

$$\lim_{a_0 \to 0} W_1(a_0; n)_{NP} = \alpha(p-q) + q \doteq q, \quad \alpha \ll 1. \quad (4.8)$$

At the other extreme of an infinitely great signal, P_n is significantly different from zero only for $x \gg 1$, while in the case of Q_n, we observe immediately from (4.6a) that the integral from $-\infty$ to $x = \log K$ is always $1-\alpha$, independent of a_0. Accordingly, for infinitely large signals the betting curve becomes

$$W_1(\infty\ ; n)_{NP} = p + (1-\alpha)q = 1 - \alpha q \doteq 1. \quad (4.9)$$

To compute the betting curve, Eq. (4.5), we introduce next an approximation based on the fact that we are primarily interested in the case of threshold reception, where the input (and output) signal-to-noise ratios are comparatively small. Since $I_0(Ra_0[2/\psi]^{\frac{1}{2}}) = 1 + R^2a_0^2/2\psi \leqq \exp(a_0^2R^2/2\psi)$, $a_0^2r^2 \ll 1$, we can express Eq. (4.4) approximately† as

$$x = \log\Lambda_n \doteq \log\mu - a_0^2 n + \frac{a_0^2}{2\psi}\sum_{l=1}^n R_l^2. \quad (4.10)$$

Making the transformation

$$y = (x - \log\mu + na_0^2)2\psi/na_0^2, \quad (4.11)$$

we can accordingly write P_n and Q_n as follows:

$$P_n(x; a_0)dx = dy\int_0^\infty \cdots \int \{\prod_{l=1}^n (R_l/\psi) \times \exp(-a_0^2 - R_l^2/2\psi) I_0(a_0R_l[2/\psi]^{\frac{1}{2}}) dR_l\} \cdot \delta(y - \sum_{l=1}^n R_l^2/n) = P_n(y; a_0)dy, \quad (4.12a)$$

$$Q_n(x; a_0)dx = dy\int_0^\infty \cdots \int \{\prod_{l=1}^n (R_l/\psi)\exp(-R_l^2/2\psi) dR_l\} \cdot \delta(y - \sum_{l=1}^n R_l^2/n) = Q_n(y)dy = P_n(y; 0)dy. \quad (4.12b)$$

In this approximation the Neyman-Pearson observer bases his decision on the distribution densities of the *mean* sum of squares of the n observations $R_1, R_2 \cdots, R_n$ of the envelope, with and without signal. The $P_n(y; a_0)$ and $Q_n(y)$ are the probability densities of the logarithm of the likelihood ratio (small-signal approx) when H_1 and H_0 are, respectively, in force. One can evaluate the

† A comparison of $I_0(2a_0r)$ with $\exp(a_0^2r^2)$ shows that the latter is very close to $I_0(2a_0r)$ for $a_0r < 1$, and even for $a_0r = 1$ is only about 15 percent greater.

integrals, Eq. (4.12), exactly,[4] to obtain

$$P_n(y;a_0)=\frac{n}{2\psi}\exp(-na_0^2-ny/2\psi)\times\left(\frac{y}{2\psi a_0^2}\right)^{(n-1)/2}I_{n-1}(2na_0[y/2\psi]^{\frac{1}{2}}),$$

$$y>0;\ =0,\ y<0, \quad (4.13a)$$

$$Q_n(y)=P_n(y;0)=\frac{n}{2\psi}\left(\frac{ny}{2\psi}\right)^{n-1}\times\frac{1}{(n-1)!}\exp(-ny/2\psi),\quad y>0;\ =0,\ y<0. \quad (4.13b)$$

This last is recognized as a χ^2 distribution density with n degrees of freedom. The first and second moments of y are found to be (*vide* Appendix III of reference 4)

$$\langle y\rangle=2\psi(1+a_0^2);\quad \langle y_0\rangle=2\psi,\quad (a_0=0);$$

and

$$\sigma^2=\langle(y-\langle y\rangle)^2\rangle=4\psi^2(2a_0^2+1)/n;\quad \sigma_0^2=4\psi^2/n. \quad (4.14)$$

The characteristic function associated with $P_n(y;a_0)$ is (Appendix III of reference 4)

$$F_y(\xi;a_0)=(1-2i\xi\psi/n)^{-n}\times\exp[2\psi i\xi a_0^2/(1-2\psi i\xi/n)]. \quad (4.15)$$

The betting curve for threshold reception becomes finally

$$W_1(a_0;n)_{NP}=p\int_{y_0}^{\infty}P_n(y;a_0)dy+q\int_0^{y_0}P_n(y;0)dy,$$

where

$$y_0=2\psi\left(1+\frac{\log(K/\mu)}{na_0^2}\right), \quad (4.16)$$

with P_n given by Eq. (4.13) in this approximation. The critical value K of the likelihood ratio is now obtained from

$$\int_0^{y_0(K)}Q_n(y)dy=1-\alpha, \quad (4.16a)$$

and the probability of a type II error $\beta_{\min}$ follows at once from Eq. (4.6b), with the help of the transformation (4.11). Figure 1 shows typical curves for P_n and Q_n under the above conditions. Although a precise evaluation of the integrals in Eq. (4.16) is not possible in closed form, for the more common cases of small a_0 and large n it is possible to show (Appendix III of reference 4) that P_n and Q_n reduce to normal probability densities with means and variances given respectively by Eq. (4.14). With these frequency functions in Eq. (4.16) we find that

$$W_1(a_0;n)_{NP}=p\int_{y_0}^{\infty}\exp[-y-\langle y\rangle()^2/2\sigma^2]dy/(2\pi\sigma^2)^{\frac{1}{2}}+q\int_0^{y_0}\exp[-(y-\langle y\rangle_0)^2/2\sigma_0^2]dy/(2\pi\sigma_0^2)^{\frac{1}{2}}$$

$$=(p/2)\{1+\Theta[a_0^2n^{\frac{1}{2}}[1-\log(K/\mu)/na_0^4]/(4a_0^2+2)^{\frac{1}{2}}]\}+q(1-\alpha), \quad (4.17)$$

and K is obtained from Eq. (4.16a), which in this approximation reduces to

$$\alpha=(1/2)\{1-\Theta[\log(K/\mu)/a_0^2(2n)^{\frac{1}{2}}]\}, \quad (4.18)$$

where Θ is the familiar error function[5]

$$\Theta(x)=(2/\pi^{\frac{1}{2}})\int_0^x e^{-v^2}dy.$$

Note that

$$\beta_{\min}=(1/2)(1-\Theta[a_0^2n^{\frac{1}{2}}/2^{\frac{1}{2}}-\log(K/\mu)/a_0^2(2n)^{\frac{1}{2}}]). \quad (4.19)$$

For moderately large values of n, P_n and Q_n can be developed in an Edgeworth series (Appendix III of reference 4); note that for the threshold cases where $a_0^2\ll 1$ we can replace $(4a_0^2+2)^{\frac{1}{2}}$ by $2^{\frac{1}{2}}$, and so the betting curve becomes a *universal* function of $a_0^2n^{\frac{1}{2}}$. A discussion of these results is given in Sec. 7 following.

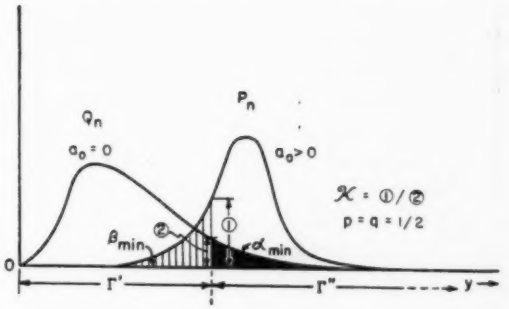


FIG. 1. The probability densities P_n, Q_n for a Neyman-Pearson observer (fixed α and constant integration time).

5. SIEGERT'S IDEAL OBSERVER

In the general case of unequal *a priori* probabilities the betting curve for Siegert's theory[3] of the Ideal observer is obtained from Eq. (4.2), except that here the critical or "signal-on" region Γ'' is distinguished from the acceptance, or "signal-off" region Γ' in such a way as to achieve a maximum of the betting curve $W_1(a_0;n)$, since the Ideal observer is by definition one who maximizes his probability of a correct decision by minimizing the *sum* of the probabilities of the type I and II errors

[4] D. Middleton, Air Force Cambridge Research Center Report No. E-5091, 1952, Appendix II.

[5] B. O. Pierce, *A Short Table of Integrals* (Ginn and Company, New York, 1929); see also, *Tables of Error Function and of Its First Twenty Derivatives* (Harvard University Press, Cambridge, 1951).

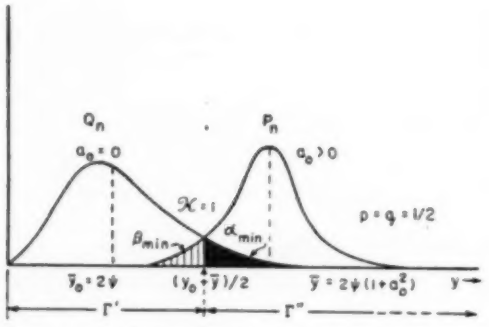


FIG. 2. The probability densities P_n and Q_n for an Ideal observer and constant integration time.

for a fixed number of observations. Under these conditions, in as much as the $W_1(R_l|a_0)$, etc., are positive, it is evident from (4.2) that $W_1(a_0;n)$ is a maximum if Λ_n exceeds unity in the "on" region and is less than unity in the "off" region. The critical value K of the likelihood ratio is therefore unity for *all* a_0. The betting curve, Eq. (4.5) accordingly reduces for the Ideal observer to

$$W_1(a_0;n)_I=p\int_0^\infty P_n(x;a_0)dx + q\int_{-\infty}^0 Q_n(x;a_0)dx, \quad (K=1), \quad (5.1)$$

and the minimized probabilities of the type I and II errors are now, respectively,

$$\alpha_{min}=1-\int_{-\infty}^0 Q_n(x;a_0)dx;$$

$$\beta_{min}=1-\int_0^\infty P_n(x;a_0)dx. \quad (5.2)$$

Observe that unlike the case of the fixed-α observer, α_{min} is not a constant quantity, independent of a_0; both α_{min} and β_{min} depend on a_0 [Eqs. (4.6a, b)]. We note further that $W_1(0;n)_I$ is explicitly

$$W_1(0;n)_I=(p-q)\times\int_0^\infty P_n(x;0)dx+q=p, \quad p>q;\ =q,\ p<q,$$

unlike $W_0(0,n)_{NP}$, which is essentially equal to q if α is small. For infinitely large signals, $W_1(\infty;n)_I$ approaches unity, since $(\alpha+\beta)_{min}$ vanishes [Eq. (4.9)]. The probability densities P_n and Q_n are then significantly different from zero only for those values of x such that $x\gg 0$ for P_n and $x\ll 0$ for Q_n; the interval of integration may accordingly be extended from $-\infty$ to $+\infty$ without changing Eq. (5.1).

As a_0 increases, the separation between the mean values of Q_n and P_n grows progressively greater, and in order to keep $K=1$ as required for the Ideal observer it is necessary to shift‡ the dividing point between the critical and acceptance regions Γ'' and Γ'. Therefore it is not sufficient, as it is in the case of the Neyman-Pearson observer, where the boundary between Γ' and Γ'' is fixed (and K varies with a_0), to consider terms in the expansion of $I_0(2a_0r)$ no higher than a_0^2. Here the "shift" in the boundary must be preserved, and to do this, terms of $0(a_0^4)$ must be included in the approximation for $x=\log\Lambda_n$. We find finally that the transformation corresponding to Eq. (4.11) becomes here

$$y=[x-\log\mu+na_0^2+(a_0^4/4)\sum_{l=1}^{n}(R_l^2/2\psi)^2]2\psi/na_0^2, \quad (5.3)$$

and since, for the case of threshold reception, it is only the dependence on a_0^2 in the additional term which is important, and not the detailed numerical factor, let us replace $\sum_l R_l^4$ by its mean value. This mean value is determined with the help of (2.16a), in as much as the signal A_0 is assumed weak compared to the noise background. One easily finds that $\sum_l\langle R_l^4\rangle$ is $8\psi^2n$ for the n independent observations. The probability densities $P_n(y;a_0)$ and $Q_n(y;a_0)$ are still given by Eqs. (4.12a, b), so that the betting curve becomes at last

$$W_1(a_0;n)_I=p\int_{y_0'}^\infty P_n(y;a_0)dy+q\int_0^{y_0'} Q_n(y)dy,$$

with

$$y_0'=2\psi(1+a_0^2/2)-[2\psi\log\mu]/na_0^2. \quad (5.4)$$

Figure 2 shows why it is necessary to include in some way the dependence of y on a_0§: the point at which K is unity occurs effectively at $(\langle y\rangle_0+\langle y\rangle)/2$ for all a_0, provided n is large. Then, as a_0 is made greater,§ this point shifts to larger values of y, and the minimized probabilities of the type I and II errors also become progressively smaller, as expected.

Making use of the Gaussian structure of P_n and Q_n when n is large (the case of threshold reception), we readily find that the betting curve, Eq. (4.16) reduces to

$$W_1(a_0;n)_I=(1/2)\{1+p\Theta[(a_0^2n^{\frac{1}{2}}+(\log\mu)/a_0^2n^{\frac{1}{2}})/2\sqrt{2}(2a_0^2+1)^{\frac{1}{2}}] +q\Theta[(a_0^2n^{\frac{1}{2}}-(\log\mu)/a_0^2n^{\frac{1}{2}})/2\sqrt{2}]\}, \quad n\gg 1. \quad (5.5$$

Usually, for the weak signals considered here $(2a_0^2+1)^{\frac{1}{2}}$ can be replaced, as before, by unity, and we see that once again the betting curve becomes in this approximation a universal function of $a_0^2n^{\frac{1}{2}}$. If the *a priori* probabilities of noise and signal are equal, Siegert's original result[6] follows immediately:

$$W_1(a_0;n)=(1/2)\{1+\Theta(a_0^2n^{\frac{1}{2}}/2^{\frac{1}{2}})\}, \quad p=q=\tfrac{1}{2}. \quad (5.6$$

‡ Note that since the critical region Γ'' changes with a_0, this test is not a uniformly most powerful one, Part I, Sec. 2a.

§ Although our original approximation, represented by the transformation, Eq. (5.3) no longer holds when a_0 becomes $0(1)$ or larger, the correct expression is still monotonically increasing with a_0, so that the essence of our argument is in no way changed.

[6] J. L. Lawson and G. E. Uhlenbeck, *Threshold Signals* (McGraw-Hill Book Company, Inc., New York), p. 172, Eq. (58a)

The probabilities of the type I and II errors here are, respectively,

$$\alpha_{\min}=(1/2)\{1-\Theta([a_0^2n^{\frac{1}{2}}-\log\mu/a_0^2n^{\frac{1}{2}}]/2^{\frac{1}{2}})\};$$
$$\beta_{\min}=(1/2)\{1-\Theta([a_0^2n^{\frac{1}{2}}+\log\mu/a_0^2n^{\frac{1}{2}}]/2^{\frac{1}{2}})\}. \quad (5.7)$$

A discussion of the betting curve and calculation of the minimum detectable signal is given in Sec. 7.

6. THE SEQUENTIAL OBSERVER

Instead of keeping the integration time ($\sim n$) fixed and adjusting either or both the probabilities of the type I and type II errors, as is the case for the Neyman-Pearson and Ideal observers, the notion of sequential sampling (Part I, Sec. 2) suggests that savings in integration time for the same accuracy are possible, or equivalently, that some reduction in the value of the minimum detectable signal can be made for the smoothing times used by the other two observers. Then, for the "off-on" experiment considered in this paper, the Sequential observer is defined as one who bases his decision on the value of the likelihood ratio $\Lambda_{n(s)}$, determined from the succession of observations $R_1, \cdots R_{n(s)}$ in any one trial: If $\Lambda_{n(s)}$ exceeds a previously set limit, a signal is judged to be present, while if $\Lambda_{n(s)}$ is less than another previously specified limit, noise alone is said to occur. These upper and lower bounds are determined [according to Eq. (2.10), Part I] solely by the magnitudes of α and β, and α and β are themselves set in advance according to the desired conditions of operation. Considered over many trials, the number of observations $n(s)$ is a random variable, fluctuating from trial to trial about the mean value $\langle n(s)\rangle=\langle n\rangle$, which may be determined from [Eq. (2.15), Part I] once α and β have been chosen.

A betting curve can be calculated for the Sequential observer in the following way. We observe first from Fig. 3 that the acceptance and critical regions are no longer contiguous, as they are for the Neyman-Pearson and Ideal observers, Figs. 1 and 2, but are separated by a zone we shall call the test interval. The sequence of observations goes on as long as $\Lambda_{n(s)}$ lies in this region, and terminates as soon as $\Lambda_{n(s)}$ exceeds K_2 or is less than K_1. (Note that the solidly shaded areas in Fig. 3 no longer represent α and β.) With this in mind we now observe that *for all trials having exactly $n(s)$ observations* the betting curve can still be represented by Eq. (4.2), except that here K_1 sets an upper bound to Γ' and K_2 puts a lower limit on Γ''. One has then

$$w_1(a_0;n(s))_S=p\int_{\log K_2}^{\infty}P_{n(s)}(x;a_0)dx$$
$$+q\int_{-\infty}^{\log K_1}Q_{n(s)}(x;a_0)dx,$$
$$\text{(exactly } n(s) \text{ observations)}. \quad (6.1)$$

where $K_1=\beta/(1-\alpha)<1$; $K_2=(1-\beta)/\alpha>1$. Here $P_{n(s)}$ and $Q_{n(s)}$ are given by Eqs. (4.3a, b) and p and q are as before the *a priori* probabilities of signal and noise. As in the case of the Ideal observer, we note here also that as a_0 becomes larger, the distribution densities $P_{n(s)}$, $Q_{n(s)}$ are more nearly centered about their mean values, which in turn become more widely separated; the test region also shifts in the direction of increasing x. Therefore, in our usual weak-signal approximation of the likelihood ratio $x=\log\Lambda_n$ [Eqs. (4.4), (4.10), and following], we may apply the same argument as was used for the Ideal observer, including terms of $0(a_0^4)$, so that the transformation here is still given by Eq. (5.3), but n is replaced by $n(s)$, which is for the moment fixed. Applying this to Eq. (6.1), we get finally

$$w_1(a_0;n(s))_S=(1/2)\{1+p\Theta[(a_0^2n(s)^{\frac{1}{2}}$$
$$-2\log(K_2/\mu)/a_0^2n(s)^{\frac{1}{2}})/2^{\frac{1}{2}}(2a_0^2+1)^{\frac{1}{2}}]$$
$$+q\Theta[(a_0^2n(s)^{\frac{1}{2}}+2\log(K_1/\mu)/$$
$$a_0^2n(s)^{\frac{1}{2}})/2^{\frac{1}{2}}]\};\ n(s) \text{ fixed}. \quad (6.2)$$

For our approximation of large (fixed) $n(s)$, i.e., weak signals with $(2a_0^2+1)^{\frac{1}{2}}\doteq 1$, we obtain again the betting curve as a universal function of $a_0^2n(s)^{\frac{1}{2}}$.

So far we have considered only those members of the ensemble of repeated trials containing exactly $n(s)$ pulses. To calculate the effective betting curve we must, of course, average Eqs. (6.1) and (6.2) over all possible values of sample size, since $n(s)$ is a random variable. Accordingly, the effective betting curve becomes

$$W_1(a_0)_S=\int_0^{\infty}W_S(n(s))w_1(a_0;n(s))_S dn(s), \quad (6.3)$$

where w_1 is given exactly by Eq. (6.1), and W_S is the probability density of the sample size. Unfortunately, W_S is not generally known, so that a precise expression for Eq. (6.3) cannot be found. However, in the usual case of threshold reception, where a_0^2 is small and consequently $n(s)$ is on the average large, we use the following

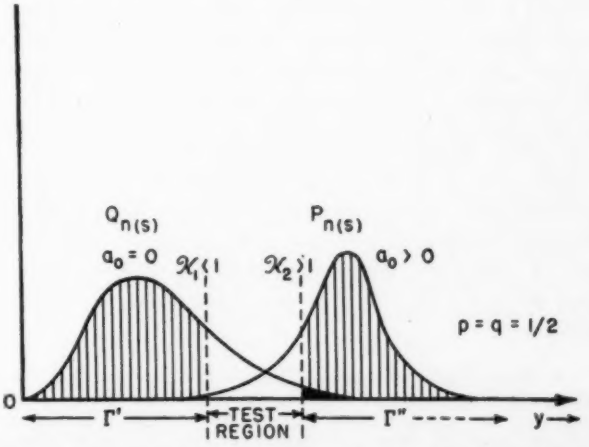


FIG. 3. The probability densities $P_{n(s)}$, $Q_{n(s)}$ for a Sequential observer showing the acceptance, test, and critical regions.

approximations to W_S in an attempt to calculate the effective betting curve Eq. (6.3):

$$W_S \doteq \delta(n(s)-\langle n\rangle), \tag{6.4a}$$

$$W_S \doteq \frac{\exp[-(n(s)-\langle n\rangle)^2/2\sigma_s^2]}{[2\pi\sigma_s^2]^{\frac{1}{2}}},$$

$$\sigma_s^2=\langle n(s)^2\rangle-\langle n(s)\rangle^2=\langle n^2\rangle-\langle n\rangle^2. \tag{6.4b}$$

The first assumes that for large $\langle n(s)\rangle$, $n(s)$ is effectively confined to values in the immediate vicinity of the mean, with negligible "tails" on either side of $\langle n\rangle$. The second is a more refined[7] model, which includes the finite spread of the distribution for $n(s)$; [note that $\lim_{\sigma_s\to 0}\exp[-(n-\langle n\rangle)^2/2\sigma_s^2]/(2\pi\sigma_s^2)^{\frac{1}{2}}\to\delta(n-\langle n\rangle)$, Eq. (6.4a)]. As we shall see presently, Figs. 6 and 8, calculations of the betting curves and minimum detectable signals on the basis of the approximate distributions (6.4) yield values too small for the former and too large for the latter.

If we use (6.4a) and apply it to (6.3), we get at once the result Eq. (6.2), where now $n(s)$ is replaced by $\langle n\rangle$. However, when the finite spread of the distribution density W_S, Eq. (6.4b), is taken into account, a more involved expression is obtained. We note from (I-16)–(I-18) of reference 4 that $\sigma_s^2 a_0^8$ is essentially independent of a_0; we see thus that in the threshold cases when $(2a_0^2+1)^{\frac{1}{2}}\doteq 1$, $a_0^2\ll 1$, the betting curve, Eq. (6.3) is a universal function of $a_0^2\langle n\rangle^{\frac{1}{2}}$, which is specifically

$$W(a_0^2\langle n\rangle^{\frac{1}{2}})_S$$

$$=(1/2)\int_0^\infty (2\pi B^2)^{-\frac{1}{2}}\exp[-(x-C)^2/2B^2]$$

$$\times\{1+p\Theta(x^{\frac{1}{2}}-A_1x^{-\frac{1}{2}})+q\Theta(x^{\frac{1}{2}}-A_2x^{-\frac{1}{2}})\}dx, \tag{6.5}$$

where

$$A_1=2^{-\frac{1}{2}}\log(K_2/\mu);\quad A_2=-2^{-\frac{1}{2}}\log(K_1/\mu);$$
$$B^2=\sigma_s^2a_0^8/64;\quad C=(a_0^2\langle n\rangle^{\frac{1}{2}})^2/8; \tag{6.6}$$

B and C are always positive, while A_1 and A_2 usually are, unless μ is very large. Unfortunately, this integral cannot be evaluated analytically in any convenient form for practical choices of α and β, so that a graphical integration is required. Tables of $W_1(a_0^2\langle n\rangle^{\frac{1}{2}})_S$ are included in Appendix V of reference 4 for a variety of values of α,β in the important case of equal *a priori* probabilities ($\mu=1$). We note also that as $a_0\to 0$, the betting curves reduce to 0 or p, respectively, if $K_2/\mu>1$ or <1, and $\mu>1$, while they become q or 0, when $\mu<1$ and $K_1/\mu>1$ or <1. For $\mu=1$, the limiting value is zero.

[7] A. Wald, *Sequential Analysis* (John Wiley and Sons, Inc., New York, 1947). Appendix A6 shows that a normal distribution of this type may be expected when $\langle n\rangle\gg 1$ when normal statistics are initially assumed. However, here one does not have normal statistics [2.16] so that (6.4a, b) can be regarded only as possible distributions, which may or may not be satisfactory representations of the true state of affairs. See Sec. 7c.

7. RESULTS AND CONCLUSIONS

It is necessary now to interpret these results for practical situations. Unless otherwise indicated, a condition of threshold reception is assumed, in which the desired signal is weak compared to the background noise, or at most comparable with it. We begin by considering first *a priori* probabilities.

A. *A priori* probabilities

Two classes of *a priori* probabilities (or probability densities) appear in the present case. These are the probability densities of the expected interference $W_1(R_1|a_0)$ and $W_1(R_1|0)$, Eqs. (2.16a, b), which are assumed known before the observer is ever asked to interpret his data. In their most general form, of course, these quantities are the joint, nth-order frequency functions $W_n(R_1,\cdots R_n|a_0)$ and $W_n(R_1,\cdots R_n|0)$, Part I, Sec. 2 [Eqs. (2.1)–(2.8)]. The second set of *a priori* probabilities are p and q, respectively, the probabilities that any one trial is likely to contain a signal, as well as noise, or noise alone. The case of chief importance occurs when $p=q=\frac{1}{2}$, where it is known for any one trial on the average that a signal and noise, or noise alone, is equally likely. The general theory discussed in Secs. 4–6 has been constructed to include cases of $p\neq q$ as well as $p=q$. More general *a priori* probabilities may be introduced, to describe known distributions of signal sources from beam-width to beam-width throughout the scanning interval, for example, but such problems are outside the scope of this paper, which aims to present only the basic theory of the statistical observation process.

B. False Alarm Time and False Rest Time

Because only a finite smoothing time is available in the course of detection, there is then a certain probability α that noise will be incorrectly reported as a signal. It is logical, therefore, to ask what is the average number of times per second, T_{FA}^{-1}, that such a false signal is indicated. In fact, this is the quantity which is usually set in practice and from which in turn is determined the value of α to be used. Specifically, T_{FA} is called the *false alarm time* or *false usage time*:[8,9] it depends on α in the following way:

$$T_{FA}=T/\alpha\langle C\rangle n\Delta t=1/\alpha\langle C\rangle[\Delta t/T_0], \tag{7.1}$$

where

$T=nT_0=$ over-all integration time,
$T_0=$ repetition period of the pulses,
$\Delta t=$ width of (time) gate about the maximum of the pulse, (7.2)
$\langle C\rangle=$ average number of crossings/sec by the envelope R of a threshold level $R_0\approx\langle R\rangle$.

[8] H. Hanse, "The Optimization and Analysis of Systems for the Detection of Pulsed Signals in Random Noise," Doctoral dissertation, Massachusetts Institute of Technology, January, 1951.

[9] M. Schwartz, "A Statistical Approach to the Automatic Search Problem," Doctoral dissertation, Harvard University, June, 1951.

Accordingly, $\Delta t/T_0$ is the relative "on"-time of the observer and $n\Delta t/T_0$ is the over-all "duty-cycle" of the observer, both of which are strongly affected by the magnitude of the gate Δt. (For the one-point theory developed here, Part I, Sec. 1, Δt is to be quite small in practice.) The threshold R_0 about which $\langle C\rangle$ is computed is somewhat indeterminate, but the mean value $\langle R\rangle$ of the noise envelope (following detection) is a suitable choice.‖ It has been shown[10] for a threshold $R_0=\langle R\rangle$ that in the case of normal random noise (after a half-wave linear detector)

$$\langle C\rangle=\rho\langle r\rangle\pi^{-\frac{1}{2}}\exp(-\langle r\rangle^2-a_0^2)I_0(2a_0\langle r\rangle)\doteq\rho\langle r\rangle\pi^{-\frac{1}{2}}\exp(-\langle r\rangle^2),\quad(a_0^2\ll 1),\quad\langle r\rangle=\langle R\rangle/(2\psi)^{\frac{1}{2}},\tag{7.3}$$

with

$$\rho^2\equiv\int_0^\infty(\omega-\omega_0)^2w(f)df\Big/\int_0^\infty w(f)df(=\psi),\quad\omega=2\pi f;\quad\omega_0=2\pi f_0,\tag{7.4}$$

where $w(f)$ is the spectral intensity of the narrow-band noise entering the second detector of the receiver and f_0 is the central frequency of this noise band. If Δf represents the bandwidth of the noise (or pulse, if the two are properly matched[2] between half-intensity points, one finds that for a Gaussian IF filter response [Table 1 of reference 10] $\langle C\rangle$ is very nearly equal to Δf. The value for the probability of the type I error to be used by the various observers is accordingly given by

$$\alpha\doteq T_0/T_{FA}\Delta t\Delta f,\tag{7.5}$$

and values of $\alpha=10^{-3}$ to 10^{-13} cover the usual practical range, the former corresponding to rather short alarm times; the latter may run into a period of days or longer.

In a similar way, a *false rest time* T_{FR} can be defined: this is the (reciprocal of the) average number of times per second an actual signal is incorrectly regarded as noise and is hence unacted upon. Analytically, one can obtain T_{FR} from Eq. (7.1) by replacing α by β, the probability of a type II error. The threshold R_0 is chosen as before to be $\langle R\rangle$, and so, like Eq. (7.5), we observe that β and T_{FR} are connected by

$$\beta\doteq T_0/T_{FR}\Delta t\Delta f.\tag{7.6}$$

However, since α and β are *not* independently chosen when the number (n) of samples is fixed beforehand [Part I, Sec. 2], α and β, and therefore T_{FA} and T_{FR} are functionally related, according to the observation process itself. This is true of the Neyman-Pearson and Ideal observers, but for the Sequential observer it is possible to select α and β independently, the integration period for a given percentage of correct decisions now varying randomly from trial to trial about the average $\langle n\rangle$.

‖ When the signal is weak there is small difference between the envelopes for noise alone and for a signal and noise.

[10] D. Middleton, J. Appl. Phys. 19, 817 (1948), Sec. 2.

C. The Betting Curves

The betting curves W_1 for the three observers have been determined in Sections 4–6. However, it is more convenient, usually, to use a "normalized" form of the betting curve, which takes into account the *a priori* probabilities p and q, i.e., the probabilities of a successful decision based on a guess as to the presence or absence of a signal. This is done by subtracting from the original, unnormalized betting curve p or q, whichever is the larger, and dividing this difference by a suitable quantity such that when $a_0\to\infty$ the betting curves become unity, as they should for an absolutely certain result. The normalized betting curves are accordingly:

$$(Neyman\text{-}Pearson)\text{—}\mathcal{P}_1(a_0^2n^{\frac{1}{2}})_{NP}=[W_1(a_0^2;n)_{NP}-(q+\alpha(p-q))]/[1-q-(\alpha p)],\tag{7.7a}$$

$$(Ideal)\text{—}\mathcal{P}_1(a_0^2n^{\frac{1}{2}})_I=[W(a_0^2;n)_I-(p\text{ or }q)]/[1-(p\text{ or }q)],\tag{7.7b}$$

$$(Sequential)\text{—}\mathcal{P}_1(a_0^2n^{\frac{1}{2}})_{Seq}=[W(a_0^2;n)_S-(p\text{ or }q)]/[1-(p\text{ or }q)].\tag{7.7c}$$

Figure 4 shows the normalized betting curves for the Neyman-Pearson observer; the Ideal observer is represented by Fig. 5, while for the Sequential observer betting curves are given in Fig. 6.

The chief feature of all these curves is the fact that $\mathcal{P}_1$ is a *universal function* of $a_0^2n^{\frac{1}{2}}$, $(n=\langle n\rangle)$, in the usual weak-signal cases, and becomes a universal function of a_0^2n, $(n=\langle n\rangle)$, for large input signal-to-noise ratios. This universal behavior breaks down, of course, in the transtition region indicated by the term "estimated" in Figs. 7, 8, 9, since then the statistics of noise alone and a signal and noise are sufficiently distinct to destroy this simple dependence on $a_0^2n^{\frac{1}{2}}$ or a_0^2n [Eqs. (4.13a)–(4.14), (4.17), for example]. In the cases of the Neyman-Pearson and Sequential observers we also observe that the smaller the choice of α, the larger is the value of $a_0^2n^{\frac{1}{2}}$ required for a given $\mathcal{P}_1$. This is reasonable, since smaller probabilities of type I errors require larger input signal-to-noise ratios for the same percentage of success, Figs. 4–6. The Ideal observer's betting curve, on the other hand, [see Fig. 5], depends only on *one* value of $\alpha=\alpha_{\min}(=\beta_{\min})$ for any single choice of $a_0^2n^{\frac{1}{2}}$, since here both $\alpha_{\min}$ and $\beta_{\min}$ are jointly determined by the process of maximizing the betting curve [Sec. 5]. For the Sequential observer, neither α nor β is adjusted by the method of observation: these quantities are decided upon beforehand. Smaller values of α *and* β then require longer integration times, i.e., large samples on the average, which is also to be expected, Fig. 6. The (unnormalized) betting curves $\mathcal{P}_1'$ of Fig. 6 for the Sequential observer include the case of a Gaussian distribution of sample size $n(s)$, Eq. (6.4b). We note

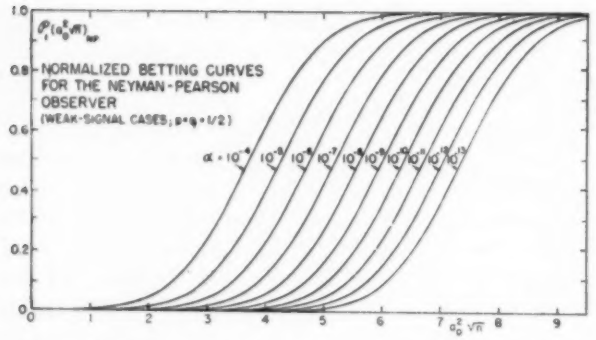


FIG. 4. Normalized betting curves $\mathcal{P}(a_0^2\langle n\rangle^{\frac{1}{2}})_{NP}$ *vs* $a_0^2\langle n\rangle^{\frac{1}{2}}$ for the Neyman-Pearson observer $[(p=q=\frac{1}{2}),\ \alpha=10^{-4}-10^{-13}]$; weak-signal cases.

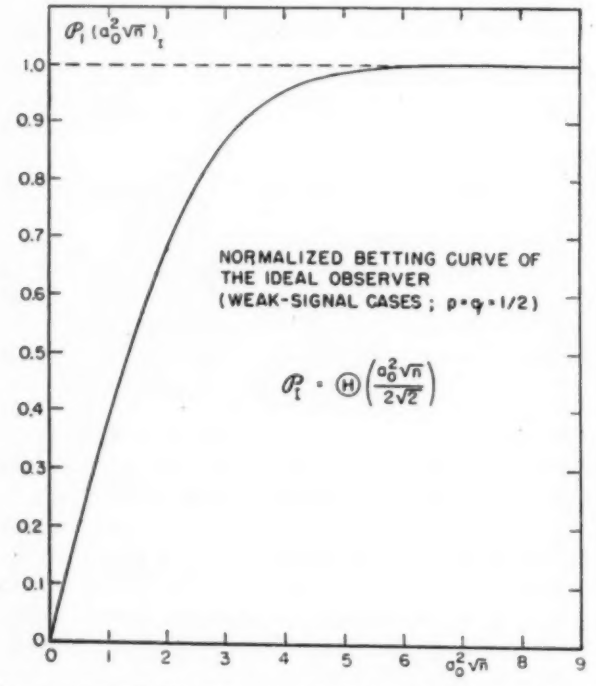


FIG. 5. Normalized betting curve $\mathcal{P}_I$ for the Ideal observer (weak-signal case) *vs* $a_0^2\langle n\rangle^{\frac{1}{2}}$; $(p=q=\frac{1}{2})$.

that in comparison with the delta-function distribution, Eq. (6.4a), the former yields betting curves which are considerably shifted in the direction of larger values of $a_0^2\langle n\rangle$, indicating that for a given choice of level of success [e.g.—$(a_0^2)_{0.90}$], larger samples on the average are required than for the former case.

D. The Minimum Detectable Signal

Once the betting curve has been found, the minimum detectable signal can be immediately specified, since by definition it is the value of a_0^2 corresponding to some arbitrarily chosen percentage of correct decisions on the average, *after the a priori probability* (p *or* q) *has been subtracted*, Eqs. (7.7a–c). A useful value is $\mathcal{P}_1=0.90$, although smaller or larger values may prove more convenient for special applications. Extremely large values (≈ 1), however, lead to too great an insensitivity: for a fixed integration time the input signal has to be excessively large compared to the noise before the system will correctly respond to a signal. In the same way, too small a choice of $\mathcal{P}_1$ for the minimum detectable signal makes the error in the decision too great for adequate operation.

Since $\mathcal{P}_1$ for each observer is a universal function of $a_0^2n^{\frac{1}{2}}$ or of a_0^2n, one can write, once the threshold defining the minimum detectable signal has been selected, say at the 90 percent level,

$$(\textit{Weak signal})\text{—}(a_0^2)_{0.9}=\frac{H_1(0.9)}{n^{\frac{1}{2}}},\quad \text{or}\quad \frac{H_1(0.9)}{\langle n\rangle^{\frac{1}{2}}}$$

(Sequential obs.), (7.8a)

$$(\textit{Strong signal})\text{—}(a_0^2)_{0.9}=\frac{H_1'(0.9)}{n},\quad \text{or}\quad \frac{H_1'(0.9)}{\langle n\rangle};\ (7.8b)$$

H and H' are called the *threshold constants;* they depend parametrically on the integration time, on the threshold, on the observation process, and depending on the type of observer, on α (and β). The behavior of the minimum detectable signal itself is shown in Fig. 7. The curves consist in all cases of two linear portions, one of slope unity for strong signals, the other of slope $\frac{1}{2}$ for the weak signals, and a curved, intermediate portion where neither the $1/n$ nor $1/n^{\frac{1}{2}}$ dependence is followed. Making α small decreases somewhat the sensitivity of the system, but this decrease is only about 2.0 db or less for the range $\alpha=10^{-4}-10^{-13}$ in the cases of the Neyman-Pearson and Sequential observers. A detailed comparison of the minimum detectable signal curves is given subsequently in Sec. F.

E. Operation of the Observer

The decisions reached by an observer of the present type require comparison of the sample's likelihood ratio

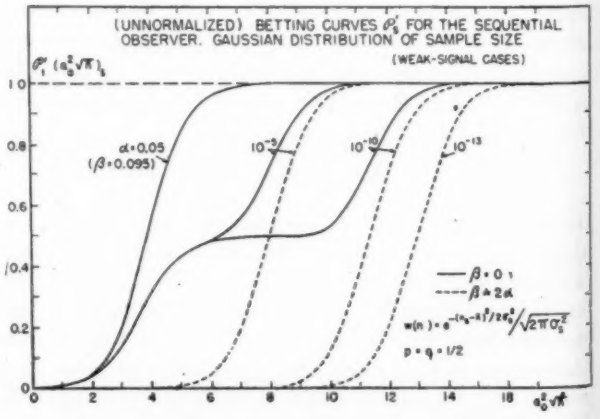


FIG. 6. Unnormalized betting curves $\mathcal{P}_s'$ for the Sequential observer, with a Gaussian distribution of sample-size (n_s) [for $p=q=\frac{1}{2}$; $\beta=0.1$ and $\beta=2\alpha$; $\alpha=5\cdot10^{-2}$; 10^{-5}, 10^{-10}, 10^{-13}].

Λ_n with some critical value of Λ_n, in the case of the Neyman-Pearson and Ideal observers, once α has been chosen, and for the Sequential observer comparison with a pair of critical values, after α and β have been selected. This procedure is common to all detection processes which are formulated as tests of a statistical hypothesis, according to the ideas outlined in Part I, Sec. 2. Note that the critical values of the likelihood ratio all contain a_0^2 as a parameter. Consequently, we must distinguish two principal conditions of operation:

(1) a_0 *is known*—when this is so we can "build into" the observer the given value of a_0. This allows us to compare the performance of one system against that of another, if we require that identical observers be employed. We have then a uniquely defined method of comparison. Furthermore, when the half-wave linear second-detector is used (Part I, Sec. 3), we recall that the three types of observer provide optimum tests under these conditions, for all a_0, since a_0 is specified in every instance.

(2) a_0 *is not known*—this is the usual situation in practice, because in actual operation the input signal-to-noise ratio is not obtainable beforehand. The n values of the detected envelope $R_1' \cdots R_n'$ are known, however, as soon as they are received. In addition, one must *choose* an a_0 in order to determine the critical likelihood ratio Λ_n against which the $\Lambda_n(R_1' \cdots R_n'; a_0)$ of a given sample or run is to be compared. The natural choice of a_0 is the value corresponding to the minimum detectable signal for some selected threshold, say the 90 percent success point on the betting curve $\mathcal{P}_1$. Then a signal weaker than this will not be indicated, and one stronger will be perceived, both decisions being correct (90 percent of the time). Again, the Neyman-Pearson, Ideal and Sequential observers yield an optimum test for $(a_0^2)_{\text{threshold}}$, but *not*, however, *for any other value of* a_0^2, *once* $(a_0^2)_{\text{threshold}}$ *has been selected.*¶ This occurs because the critical region Γ'' is not independent of the threshold, as explained in Sec. 2: we do not here have uniformly most powerful tests.** In other words, while signals above the threshold are indicated, as required, and decisions made with *less* than the percent error set by our choice of $(a_0^2)_{\text{threshold}}$, these decisions are not made with the least error compatible with the actual magnitude of a_0^2. When we are concerned only with the presence or absence of a signal, this is certainly not serious: we get our answer in the fixed integration time set for the trial at least with no greater than the maximum decision error (10 percent if $(a_0^2)_{0.90}$ is chosen). However, when the signal is stronger than $(a_0^2)_{\text{threshold}}$, we may wish this answer (at the maximum error of $(a_0^2)_{\text{threshold}}$) in a shorter time; clearly some loss of efficiency results from the fact that the test is no longer an optimum one for $(a_0^2)_{\text{actual}} \neq (a_0^2)_{\text{threshold}}$ for these stronger signals. The Sequential observer overcomes this difficulty, because such an observer adjusts the sample size according to signal strength. Stronger signals accordingly require a shorter smoothing time, but in any case, since a_0 is not known, some threshold value of it must (as for the other observers) be built into the observer.

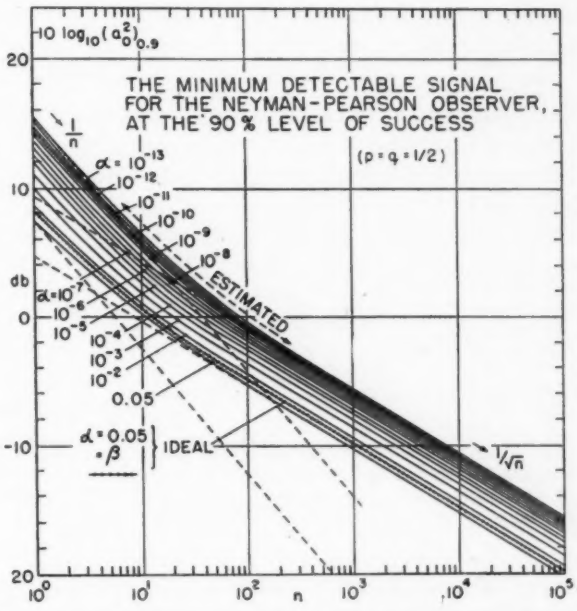


FIG. 7. The minimum detectable signal for the Neyman-Pearson and Ideal observers, as a function of integration time (n) at the 90 percent level of success. [$\alpha = 0.05 - 10^{-13}$; $p = q = \frac{1}{2}$.]

F. Comparison of the Neyman-Pearson, Ideal, and Sequential Observers

Having specified a series of optimum tests for our problem of determining the presence or absence of a pulsed carrier in noise, we must now determine which of these is the best, when this is a meaningful question. The comparison, of course, must be made on the basis of identical tasks for identical purposes: it makes little sense, for example, to test system A against system B, when the two are designed to fulfill different duties.

For the moment we confine our attention solely to the problem of monitoring a single range interval. Then *the best observer* is here defined as *the one which, for a given integration time, and a given value of the type I error, requires the smallest minimum detectable signal for a previously selected level of success on the betting curve* (i.e., the smallest value of, say, $(a_0^2)_{0.9}$ for a given $n(=\langle n \rangle)$

¶ The tests in (1) above are optimal, because $(a_0^2)_{\text{threshold}}$ is adjusted for each run, and represents the value of the actual input signal-to-noise ratio.

** The fact that our tests are not independent of the threshold, depends, of course, directly on the specific character of the *a priori* probability densities $W_n(R_1 \cdots R_n | a_0)$ and $W_n(R_1 \cdots R_n | 0)$ of the interfering noise (Sec. 2). For some other statistics it is possible that the tests may be uniformly most powerful ones, but this does not appear likely, since very few physically useful random processes are known for which this can occur. We also observe that α in all (weak-signal) cases is always determined from the false-alarm time (B), not β, inasmuch as α can be determined from the *known* distribution $W_n(R_1 \cdots R_n | 0)$, whereas β requires knowledge of the correct a_0 to be inserted in $W_n(R_1 \cdots R_n | a_0)$, a datum which is not generally available.

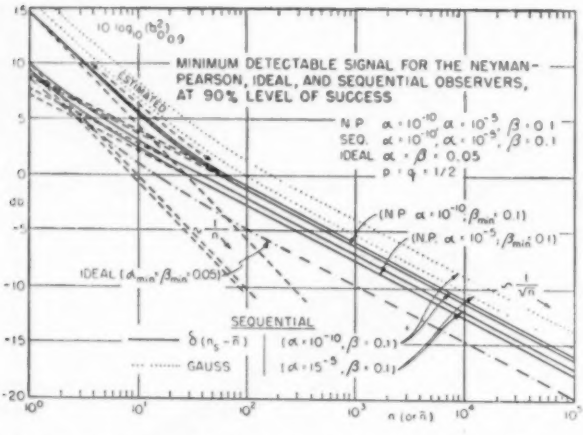

MINIMUM DETECTABLE SIGNAL FOR THE NEYMAN-PEARSON, IDEAL, AND SEQUENTIAL OBSERVERS, AT 90% LEVEL OF SUCCESS
ESTIMATED
N.P. α = 10⁻¹⁰, α = 10⁻⁵, β = 0.1
SEQ. α = 10⁻¹⁰, α = 10⁻⁵, β = 0.1
IDEAL α = β = 0.05
P = q = 1/2
db
(N.P. α = 10⁻¹⁰, β_min = 0.1)
(N.P. α = 10⁻⁵, β_min = 0.1)
IDEAL (α_min, β_min = 0.05)
1/√n
SEQUENTIAL
δ(n_s − n̄)
(α = 10⁻¹⁰, β = 0.1)
(α = 10⁻⁵, β = 0.1)
GAUSS
n (or n̄)

FIG. 8. Minimum detectable signal for the Neyman-Pearson, Ideal, and Sequential observers. [$\alpha=10^{-5}$, 10^{-10}; $\beta=0.1$ for $N-P$ and Sequential observers; $\alpha_{\min}=\beta_{\min}=0.05$ for Ideal observers; all at 90 percent level of success.]

and α). We cannot always simultaneously require the probability of the type II error β to be identical for each observer, since β is frequently determined (i.e., minimized) by the observation process itself. It is found, for example [Fig. 8.21 of reference 4], in the weak-signal cases ($a_0^2<0$ db), that with equal probabilities of type I errors the Neyman-Pearson observer is best at the 90 percent level of success, by about 0.5 db over the Ideal observer (which, however, yields a smaller probability of a type II error), by about 1.5 db over the approximate results for the Sequential observer with a delta-function distribution of sample-size ($n_s=\langle n\rangle$), and by almost 3.0 db over the approximate results for the Sequential observer with a Gaussian distribution of sample-size.

A not unexpected difficulty should be pointed out when the Ideal observer is compared with either the Neyman-Pearson or Sequential observers. For once the level of success (here taken to be the 90 percent point on the *normalized* betting curves, Eqs. (7.7), has been selected for the Ideal observer, the probability of the type I error is determined, unlike the case of the other observers, where α may be chosen, quite independent of the level of success. With $(a_0^2)_{0.9}$, $\alpha_{\min}$ is 0.05 for the Ideal observer, but for the false alarm times of any practical use α is from Eq. (7.5) required to be much smaller, usually $0(10^{-4})$ or less. Such values of α, in turn, set a much higher level of success in the Ideal observer: for example, the value $\alpha=10^{-5}$ corresponds here to the 99.998 percent point on the betting curve. This is usually too great an accuracy of decision for ordinary use, being about 1.5–2.0 db above the 90 percent point for any given sample size $n(=\langle n\rangle)$. Figure 9 gives the minimum detectable signal as a function of the number of pulses integrated under this condition of a much higher level of success (x). Again, the same order of superiority is preserved, as in the $x(=90$ percent) case, [Fig. 8.21 of reference 4]. The Sequential observer, with†† $w_s=\delta(n_s-\langle n\rangle)$, needs the largest input, the Neyman-Pearson the least (by about 2.5 db) for a given $n(=\langle n\rangle)$, with the Ideal observer just slightly inferior to the Neyman-Pearson.‡‡ These smaller probabilities of type I errors, as expected, demand larger values of the threshold constants, since a more accurate decision is required in the allotted time. We find also that the Neyman-Pearson observer appears to suffer the least increase in the minimum detectable signal as the percent level of success (x) is raised, while the Ideal observer is least stable in this respect, exhibiting the largest shift in $(a_0^2)_x$ as x goes from 90 percent to 99.99$\cdots$98 percent.

Finally, if we consider only the Sequential and Neyman-Pearson observers we can compare them at the more practical (90 percent) level of success, with the same α *and* β, if β_{seq} is set equal to $\beta_{\min}$ of the Neyman-Pearson test; α now is some sufficiently small value determined in the usual way by the false alarm time. The comparison follows from the curves of Fig. 8, where the Neyman-Pearson gives an apparent superiority over the Sequential observer, as indicated, by 0.5 and 2.5 db for $\alpha=10^{-5}$, and by about the same amounts for $\alpha=10^{-10}$.

From the preceding discussion the following general observations can be made:

(1) The Neyman-Pearson observer is better than the Ideal observer by about 0.5 db. It is more stable, i.e., less sensitive to changes in threshold. In fact, this observer in the weak-signal cases is essentially independent of threshold,§§ and so is effectively a uniformly most powerful observer. It is this feature, as well as the low value of its threshold constant, which makes this observer particularly attractive in many practical situations.

(2) The Ideal observer is next in preference from the standpoint of a low minimum detectable signal (for fixed integration time). However, it is the least stable of the three, i.e., it gives the largest departures from optimum operation for inputs larger than the threshold value at the 90 percent level, which is built into the likelihood ratio Λ_n. It is also somewhat more sensitive than the Neyman-Pearson observer to changes in input level. The chief drawback is the high level of success (x) required for practical values of the probability of a type I error. Comparing the curves in Figs. 8 and 9 for $\alpha=10^{-5}$, for instance, we see that the Ideal is about 1.7 db worse at $x=0.99998$ than the Neyman-Pearson observer is at $x=0.90$.

†† Curves for the Sequential observer with a Gaussian distribution of n_s are not shown here, but it is estimated that they should lie about 1.7–2.0 db above the Sequential observer's curves for $w_s=\delta(n_s-\langle n\rangle)$.

‡‡ The difference between the results for the Neyman-Pearson and Sequential observer with $w_s=\delta(n_s-\langle n\rangle)$ is due specifically to the approximation (2.10) for discrete sampling, and the approximations inherent in the calculation of $\langle n\rangle$, (2.14), (2.15). For $n_s=\langle n\rangle=n_{NP}$ one should, of course, obtain precisely the Neyman-Pearson result.

§§ In theory (Sec. 4), only for $a_0=0$ is the Neyman-Pearson observer strictly independent of the input signal level.

(3) For the approximate distributions of sample size considered in this paper, the Sequential observer requires a higher minimum detectable signal (0.5 db for $w_s = \delta(n_s - \langle n \rangle)$, and 3.0 db for a gauss distribution of n_s) than the Neyman-Pearson observer. However, we know that the Sequential test is never *less* effective than the corresponding Neyman-Pearson procedure, since the latter test can always be regarded as a special case of sequential sampling in which the test terminates after exactly $n = \langle n \rangle$ samples. The apparent inferiority of the sequential approach here is due to the necessarily approximate nature of the distributions of sample size, the true ones being as yet undiscovered. A *lower* bound on the errors thus introduced in the minimum detectable signal is accordingly provided by the extent to which the Sequential observer appears worse than the Neyman-Pearson observer for equal probabilities of the type I and II errors α, β. For this reason Fig. 9 and the Sequential curves of Figs. 8 and 9 have been included in the present discussion. A complete theory of the Sequential observer remains to be developed.

G. General Remarks on the Statistical Theory of the Observation Process

There are two main parts to the problem of observing the signal in noise: the first is the transfer of the incoming message (and noise) from rf to video frequencies without either losing the available information or inserting false data in the process of observation. The second is the processing or interpretation of the available information. In the former instance we are concerned only with making available, without prejudice, the legitimate data; in the latter, we are asked to make a decision on the basis of this information, the nature and method of this process being set wholly or in part by the ultimate use to which the incoming data are to be put. In statistics, the former is recognized as a problem in the "estimation of parameters," while the latter, as we have already seen in Part I, Sections 1 and 2, belongs to the theory of testing statistical hypotheses.

As an illustration, let us consider the question of locating the signal pulse in a range interval (large compared to the pulse width), it being known *a priori* that a signal of a given shape is present. As Woodward has pointed out,[11] the only proper formulation of the problem is to ask for the *probability distribution* of the pulse's position (measured from some distinguishing point, usually at the maximum amplitude), *not its exact location*. As soon as we force ourselves to an exact answer, on the basis of necessarily incomplete knowledge, we introduce spurious data and thus destroy some of the available information. We deny to a certain extent the basically statistical character of the whole process, or in other words, the attempt at an exact answer is a kind of cheating, the penalty for which is loss of information. To handle the problem, Woodward observes that only three probability distributions are needed:

(1) *the a priori distribution* (density), representing the observer's initial state of knowledge. This must include p, q, and the *form* (W_n) of the (n-fold) probability density of signal and noise. The value of the statistical parameter, here pulse location (or in our problem of the "off-on" experiment, the input signal-to-noise ratio a_0), is of course not known beforehand.

(2) *the* (maximum) *likelihood distribution* (density), which is determined by the data only, and requires *no* knowledge of the statistical parameters. The data are inserted into the *a priori* expression for W_n, and a value of the statistical parameter *chosen* which maximizes this likelihood function.[12] These data represent the addition to the observer's previous state of knowledge.

(3) *the a posteriori distribution* (density), which after suitable normalization, gives the observer's final state of knowledge; it is the (normalized) product of the *a priori* and likelihood distributions.

It is this *a posteriori* distribution which is the required answer to our original question, and which contains all the available data. In appearance, the distribution extends over the allowed range interval, is always positive, and exhibits a number of peaks, one of which is associated with the postulated signal, whose location is to be estimated. We stress the fact that this distribution curve is *not* the same as the original data "trace" in the video, but is a proper probability distribution density.

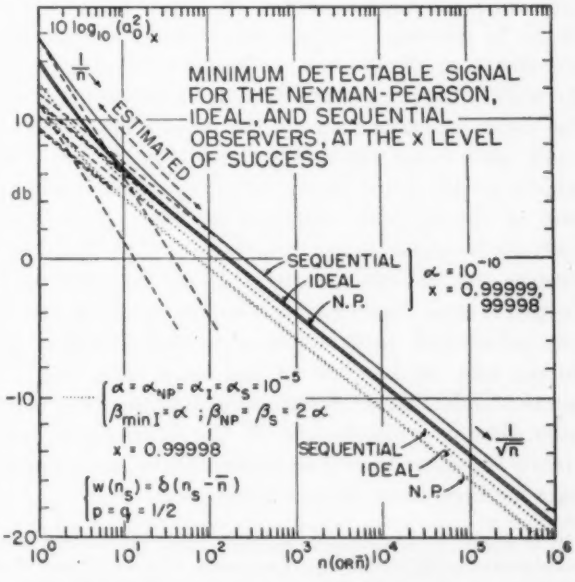


FIG. 9. Minimum detectable signal for the Neyman-Pearson, Ideal, and Sequential observers, at the x level of success. [$\alpha = 10^{-5}$, 10^{-10}; $x = 0.99998$, $0.9999\cdots99998$; $\beta_{NP} = \beta_{seq.} = 2\alpha$; $\beta_{\min I} = \alpha_I = \alpha$.]

[11] P. M. Woodward, "The theory of observation," Telecommunications Research Establishment Journal (England), 32, (January, 1951).

[12] A. Mood, *Introduction to the Theory of Statistics*, McGraw-Hill Book Company, Inc., New York (1950), p. 257 and following.

It is not an enhanced signal in noise, either, but a distribution which no amount of guessing can enhance. Furthermore, the nonlinear device by which the envelope information is transferred from the high frequency- to the video stage is determined alone by the form of the *a priori* statistics of noise and signal; here, as in Section 3, the optimum detector is the log I_0 device. Note that nothing has as yet been said about the use of these data in making a decision; for the moment our problem is simply the estimation of a statistical parameter (pulse position). To obtain a decision, the next step is to relate the *a posteriori distribution* in an appropriate way to the probabilities of type I and II errors. This may be done using a maximum likelihood test of the hypothesis of noise *vs* signal and noise, or by replacing the step (2) above by the procedures outlined for the Neyman-Pearson, sequential tests, etc. In any case, a statistical test necessarily introduces the concept of the type I and II errors, the natural and convenient use of which is made through the betting curve,‖‖ by which in turn the minimum detectable signal may be defined.

There is no unique way of deciding which type of test, likelihood ratio, Neyman-Pearson, etc., is best. One must compare them from the point of view of the same desired result. Furthermore, in the light of the above, we see that they differ only in their methods of reaching a decision; the processing of the data for a given purpose in one instance may be superior to that of another treatment. Although we have not examined it here, the maximum likelihood observer is expected to be equivalent to the Neyman-Pearson in performance. Both require knowledge of the same *a priori* distributions, both are relatively insensitive to threshold ($\sim a_0$), the former by definition, the latter essentially so because of the structure of the test and the character of the statistics in the present problem, and both are capable of the same generalization to the m-position observer, which may cover the entire range. Finally, both are subject to the same limitations: one must know the *form* of the *a priori* distribution. For a practicable analysis in most cases this means that successive samples are to be uncorrelated, so that the usual statistical tests, which are all based on *random* sampling, may be applied. In principle, correlated samples, e.g., clutter, can be handled in the same way, but the practical difficulties of determining, for example, the nth order Gaussian distribution of the envelope of noise when n is greater than 2 seem prohibitive, and approximate methods must then be attempted.

‖‖ It should be emphasized, moreover, that the betting curve introduces no additional information, real or spurious, under these conditions; it simply provides a natural connection between the information and the decision to be made on the basis of this information. Practically, of course, a little information is lost (the so-called topological error) (reference 11) when continuous sampling is replaced by discrete, i.e., when a 1-point theory is used, where really a (infinitely) many-point theory is ideally demanded. This information loss may be ignored, however, if the gate's duration is equal to or less than the pulse width.

ACKNOWLEDGMENTS

The author wishes to express his gratitude to Mrs. Virginia Johnson, who carried out most of the calculations in this paper. He also wants to thank Professor J. B. Wiesner (Massachusetts Institute of Technology), Professor H. Brooks (Harvard University), and Mr. William H. Huggins (Air Force Cambridge Research Center) for their interest in the development of this work. And finally, he wishes to convey his special appreciation to Professor Gottfried Noether (Boston University) for his helpful criticisms of the statistical features of this and the preceding paper.

APPENDIX. THE STRONG-SIGNAL CASE

Although the detection problem is no longer critical when the signal is strong [0(10 db)] compared to the interfering noise, it is of some interest to extend the curves for $(a_0^2)_{\min}$, the minimum detectable signal, Figs. 7–9, into the region of comparatively short integration time s$[1 \leq (n, \langle n \rangle) \leq 0(10)]$ and to determine explicitly how $(a_0^2)_{\min}$ depends on n (or $\langle n \rangle$) under these conditions. Unlike the case of threshold signals, where the relatively weak signal is suppressed by the noise,[2,13] so that the quadratic approximation, Eq. (4.10) applies, we have instead somewhat the reverse situation—the noise is now suppressed by the signal, and we can write our "decision" approximation, Eq. (4.10) as

$$x = \log \Lambda_n \gtrless \log \mu - n a_0^2 + 2 a_0 \sum_{l=1}^{n} R_l/(2\psi)^{\frac{1}{2}}. \quad \text{(A.1)}$$

Here the observer adds the pulses, basing his decision on the mean value of this sum, in place of the threshold criterion, which is based on the mean value of the sum of the *squares* of the noise and signal and noise envelopes. The transformation, Eq. (4.11) accordingly becomes

$$y = (x - \log \mu + n a_0^2)[a_0 n (2/\psi)^{\frac{1}{2}}]^{-1} = \sum_{l=1}^{n} R_l/n, \quad \text{(A.2)}$$

while P_n and Q_n, Eqs. (4.12a, b) are now obtained from

$$P_n(y; a_0) = \int_{-\infty}^{\infty} e^{i\xi y} (d\xi/2\pi) \left\{ \int_0^{\infty} (R/\psi) \right.$$

$$\left. \times \exp(-a_0^2 - R^2/2\psi - i\xi R/n) I_0(a_0 R[2/\psi]^{\frac{1}{2}}) dR \right\}^n;$$

$$Q_n(y) = P_n(y; 0). \quad \text{(A.3)}$$

As shown in Appendix IV, reference 4, approximate forms of P_n and Q_n, useful in the ranges $[1 \leq n \leq 0(10)]$,

[13] D. Middleton, Proc. Inst. Radio Engrs. 36, 1467 (1948).

($a_0^2 > 3$ or 4), are

$$P_n(y; a_0) \simeq [2\pi\psi/n]^{-\frac{1}{2}} \times \exp\{-[y - a_0(2\psi)^{\frac{1}{2}}]^2/2\psi/n\} - [(n/2)^{\frac{1}{2}}(\psi/n)^{-\frac{1}{2}}/4a_0]\phi^{(1)}(\lambda) + [(n-2)/32a_0^2(\psi/n)^{\frac{1}{2}}]\phi^{(2)}(\lambda) + 0(a_0^{-3}); \quad \lambda = [y - a_0(2\psi)^{\frac{1}{2}}]/(\psi/n)^{\frac{1}{2}}; \quad \text{(A.4a)}$$

$$Q_n(y) \simeq (2\pi\sigma_y^2)^{-\frac{1}{2}} \exp[-(y - \langle y\rangle_0)^2/2\sigma_y^2] + (C_3/6\sigma_y)\phi^{(3)}(\lambda') + (C_4/24\sigma_y)\phi^{(4)}(\lambda') + \cdots \quad \lambda' = (y - \langle y_0\rangle)/\sigma_y, \quad \text{(A.4b)}$$

where $C_3 = 0.631/(n)^{\frac{1}{2}}$ and $C_4 = 0.246/n$, and $\langle y_0\rangle = (\pi\psi/2)^{\frac{1}{2}}$; $\sigma_y^2 = \psi(2 - \pi/2)/n$.

In the first approximation the betting curve for the Neyman-Pearson observer is accordingly a universal function of $a_0(n)^{\frac{1}{2}}$ (if a_0 is large enough and n sufficiently small). The betting curve for the other observers, while not strictly universal functions of $a_0(n)^{\frac{1}{2}}$ or $a_0\langle n\rangle^{\frac{1}{2}}$ may be treated essentially as such in this approximation. Thus the minimum detectable signal (power) obeys a n^{-1}-law if the signal is strong and an $n^{-\frac{1}{2}}$-law in the threshold cases, as has been shown by other methods[14] (which, however, do not give the threshold constant, or the dependence on α and β).

[14] Reference 3, Sec. 7.4, Eqs. (38) and (42).

The curves for a minimum detectable signal include an interval from $n = 1$ to $n = 0(10^2)$, which consists of two rather loosely defined regions (1), the strong signal interval, for which a_0^2 is 0(10 db) and $n = 1$ to $n = 0(10)$, and (2) a threshold interval in which a_0^2 drops from about 10 db to about 0 db or less, with $n = 0(10) - 0(10^2)$. The curves for the interval $n = 1$ to $n = 0(10^2)$ are approximate, as the relations derived in Section 7 and Appendix IV of reference 4 in many cases need to be extended for accurate results. However, it is possible to get acceptable values for $n = 1$, except for the Ideal observer. Here the approximation (7.1) of the likelihood ratio Λ_n is no longer good in the strong-signal, or threshold intervals, so that a direct (and extremely laborious) calculation using the complete expression for Λ_n, involving the Bessel function, must be employed. Actually, from the behavior at $n = 1$, from the fact that when n exceeds $0(10^2)$ the accuracy [as Siegert (3) has shown] is within 10 percent or better of the true value, and from the fact that the slopes in the strong-and weak-signal regions are respectively unity and $\frac{1}{2}$ (on a db *vs* log-scale), we can estimate the desired minimum detectable signals within about 0.5 db, erring somewhat on the side of too low values. In the case of the Sequential observer, when the signal level is high, we have made the rather drastic assumption that sample size $n(s)$ has a delta-function distribution.

Use of Tschebyscheff-Polynomial Operators in the Numerical Solution of Boundary-Value Problems

GEORGE SHORTLEY
Operations Research Office, The Johns Hopkins University, Chevy Chase, Maryland
(Received August 11, 1952)

This paper concerns the numerical solution, in two or more dimensions, of boundary-value problems arising from linear partial differential equations, of which Laplace's and Poisson's equations furnish simple examples. Only techniques suitable for automatic computing machines are considered. The only method previously applicable to automatic machines is a straightforward iteration of an appropriate difference operator applied to an assumed trial solution on a network of points; as has been repeatedly pointed out, the rate of convergence of this iterative procedure is inordinately slow when the number of net points is large. The present paper shows how the employment of a Tschebyscheff polynomial of this same difference operator can cut the amount of work required in obtaining a solution by a very large factor—a factor of the order of magnitude of $\sqrt{N}$ in the two-dimensional case, $\sqrt[3]{N}$ in the three-dimensional case, where N is the number of net points. This method is an outgrowth of that developed by Flanders and Shortley [J. Appl. Phys. **21**, 1326 (1950)] for employment of such Tschebyscheff-polynomial operators in the numerical solution of *eigenvalue* problems. The method is illustrated by a simple example involving Poisson's equation.

INTRODUCTION

THIS paper considers the numerical solution of boundary-value problems arising from elliptic linear partial differential equations, of which Laplace's and Poisson's equations furnish simple examples.

We are interested in problems in two- or three-dimensional regions of sufficient complexity so that hundreds or thousands of points are required to obtain adequate accuracy in an approximating solution on a network of points. In such cases, use of high speed computing machinery is justified or required, and some of the short cuts that have been developed for hand methods are inapplicable.[1]

The only method previously applicable to automatic-machine techniques is a straightforward iteration of an appropriate difference operator applied to an assumed trial solution on the net; as has been repeatedly pointed out, the rate of convergence of this iterative procedure is inordinately slow when the number of net points is large.

The present paper shows how the employment of a Tschebyscheff polynomial of this same difference operator can cut the amount of work required in obtaining a solution by a very large factor—a factor of the order of magnitude of $\sqrt{N}$ in the two-dimensional case, $\sqrt[3]{N}$ in the three-dimensional case, where N is the number of net points. This idea is an outgrowth of the employment, by Flanders and Shortley,[2] of such Tschebyscheff-polynomial operators in the numerical solution of *eigenvalue* problems.

[1] For a review of previous work on numerical methods applicable to boundary-value problems in two or three dimensions, see Thomas J. Higgins, *Numerical Methods of Analysis in Engineering* (The Macmillan Company, New York, 1949), Chapter 10. This chapter contains an extensive bibliography.

[2] Donald A. Flanders and George Shortley, J. Appl. Phys. **21**, 1326 (1950), which see for many details not repeated in the present paper. Two minor errors have been detected in this reference: Equation (11) should have the factor 2^{p-1} inserted on the right-hand side, and the entry $m=6$, $r=0.9$ in Table I should read 0.03932.

THE DIFFERENCE EQUATION

While the method described below is applicable to boundary-value problems based on a wide variety of homogeneous and inhomogeneous elliptic linear partial differential equations, we can best illustrate the method by describing a simple example. Poisson's equation furnishes the simplest nonhomogeneous equation (of which Laplace's equation is a special, homogeneous case), and we shall use this equation for illustration.

For simplicity, we shall consider a rectangular two-dimensional region, although the method is readily generalized to more dimensions, and to irregular boundaries.[3] We shall write Poisson's equation as

$$\Delta u+\alpha(x, y)=0, \tag{1}$$

where $\alpha(x, y)$ is a given function of position. The boundary-value problem requires the solution of (1) that has given specified values of u along the boundary of the two-dimensional region.

We replace the continuum by a square net of points of spacing h, and the functions u and α by their values u_{ij} and α_{ij} at the net points. Equation (1) is then approximated by the difference equation

$$\omega u_{ij}+\tfrac{1}{4}h^2\alpha_{ij}=u_{ij}, \tag{2}$$

where ω is the operator that averages the four neighboring values:

$$\omega u_{ij}\equiv\tfrac{1}{4}(u_{i+1,j}+u_{i-1,j}+u_{i,j+1}+u_{i,j-1}). \tag{3}$$

The solution of (2) approximates that of (1) more and more closely as h is made smaller.[1]

THE ITERATION PROCEDURE

In the iteration procedure one starts with a "trial" function v_{ij} having the correct values at the boundary

[3] The "improvement formula" for a point near an irregular boundary is given in Weller, Shortley, and Fried, J. Appl. Phys. **11**, 283 (1940).

points. Equation (2) is then used repeatedly as an "improvement formula" to give a series of functions that eventually converges to the true solution u_{ij}. Let us define the inhomogeneous operator O by the relation

$$Ov_{ij} \equiv \omega v_{ij} + \tfrac{1}{4}h^2 a_{ij}. \tag{4}$$

Then the sequence v_{ij}, Ov_{ij}, O^2v_{ij}, O^3v_{ij}, $\cdots$ converges to u_{ij}, as we can readily prove:

Let

$$v_{ij} = u_{ij} + \phi_{ij}, \tag{5}$$

where u is the true solution of (2) and ϕ is the "error" in the trial function v. The error ϕ will vanish at the boundary points. Then (omitting subscripts),

$$\begin{aligned} Ov &= \omega u + \omega\phi + \tfrac{1}{4}h^2\alpha \\ &= u + \omega\phi, \\ O^2v &= u + \omega^2\phi, \\ O^3v &= u + \omega^3\phi, \cdots. \end{aligned} \tag{6}$$

The sequence $\omega\phi$, $\omega^2\phi$, $\omega^3\phi$, $\cdots$ converges to zero because ϕ has zero boundary values and ω is of the nature of an averaging operator; hence the sequence in (6) converges to u. We can expand[2] ϕ in terms of the eigenvectors $\phi^{(1)}$, $\phi^{(2)}$, $\cdots$, $\phi^{(N)}$ of the operator ω. These eigenvectors satisfy the equation

$$\omega\phi^{(k)} = \lambda_k\phi^{(k)}, \tag{7}$$

where the eigenvalues λ_k have the property

$$1 > \lambda_1 > \lambda_2 \geqq \lambda_3 \geqq \cdots > \lambda_N > -1, \tag{8}$$

and N is the number of internal points in the net.

If we write this expansion as

$$\phi = \sum c_k\phi^{(k)}, \tag{9}$$

then

$$\omega^n\phi = \sum c_k\lambda_k{}^n\phi^{(k)}, \tag{10}$$

and $\omega^n\phi$ is seen to converge to zero because the λ's are all less than unity in absolute value. The rate of convergence is determined by the size of λ_1 and λ_N, which are generally close to unity in absolute value. In a rectangular region of $p \times q$ net points,

$$\begin{aligned} \lambda_1 = -\lambda_N &= \tfrac{1}{2}[\cos(\pi/p) + \cos(\pi/q)] \\ &\approx 1 - \tfrac{1}{4}(\pi^2/p^2) - \tfrac{1}{4}(\pi^2/q^2). \end{aligned} \tag{11}$$

If we desire the largest coefficient in (10) to be a small number ϵ, we must iterate a number of times n such that

$$\lambda_1{}^n = \epsilon \quad \text{or} \quad n = \log\epsilon/\log\lambda_1. \tag{12}$$

For a square net of $p \times p = N$ points, (11) becomes $1 - \lambda_1 = \frac{1}{2}\pi^2/N$, a very small number if N is large. Then $\log\lambda_1 \approx \lambda_1 - 1 = -\frac{1}{2}\pi^2/N$, and (12) becomes

$$n = 2(N/\pi^2)(-\log\epsilon), \tag{13}$$

a very large number.

EMPLOYMENT OF THE TSCHEBYSCHEFF-POLYNOMIAL OPERATOR

Going back to (6), we see that if we apply to the trial function v a polynomial $P(O)$ in the operator O, we obtain

$$P(O)v = P(1)u + P(\omega)\phi, \tag{14}$$

where $P(1)$ is the value of the polynomial at argument unity. If we require that

$$P(1) = 1, \tag{15}$$

and use the result (10), we obtain

$$P(O)v = u + \sum c_k P(\lambda_k)\phi^{(k)}. \tag{16}$$

We now see that there is a "best" polynomial of order m to apply to v in order to obtain u, namely, that polynomial satisfying (15) that has the least maximum absolute value at all eigenvalues λ_k. Since the number N of eigenvalues is large compared to the order m of polynomial that it is feasible to use, we must apply this criterion in the following form: *The best polynomial $S_m(O)$ of order m to use in* (16) *is that satisfying* (15) *and having the least maximum absolute value throughout the range from* $\lambda_N = -\lambda_1$ *to* $+\lambda_1$. Flanders and Shortley[2] show that this best polynomial is

$$S_m(O) = T_m(O/\lambda_1)/T_m(1/\lambda_1), \tag{17}$$

where $T_m(\mu)$ is the Tschebyscheff polynomial obtained by expanding

$$\begin{aligned} T_m(\mu) &= \cos(m \arccos\mu), \quad (\mu \leqslant 1) \\ T_m(\mu) &= \cosh(m \operatorname{arccosh}\mu), \quad (\mu \geqslant 1) \end{aligned} \tag{18}$$

in powers of μ. The denominator in (17) is designed to make $S_m(\lambda) = 1$ at $\lambda = 1$.

The technique of employing such a polynomial is then, in principle, the following: Compute the series of functions v, Ov, O^2v, $\cdots$, O^mv, and then, with an estimated value of λ_1 (preferably slightly greater rather than slightly less than the correct value), form the linear combination $S_m(O)v$. This linear combination will have all coefficients of $\phi^{(k)}$ in (16) reduced to a small (and computable) fraction of their original values, since

$$S_m(O)v = u + \sum c_k S_m(\lambda_k)\phi^{(k)}. \tag{19}$$

COMPARISON OF THE METHODS

Because the Tschebyscheff polynomial $T_m(\mu)$ has unity (repeated $m+1$ times) as its greatest absolute value in the range $-1 \leqslant \mu \leqslant +1$, the function $S_m(\lambda)$ has $1/T_m(1/\lambda_1)$ as its greatest absolute value in the range $-\lambda_1 \leqslant \lambda \leqslant +\lambda_1$. The value $1/T_m(1/\lambda_1)$ can be computed accurately from (18), and is tabulated in Table I of the reference in footnote 2. All coefficients c_k in the last term of (19) are multiplied by a quantity at least as small as $1/T_m(1/\lambda_1)$.

Reference 2 did not estimate the saving in labor resulting from the use of the Tschebyscheff polynomial. In order to make a comparison with the iteration pro-

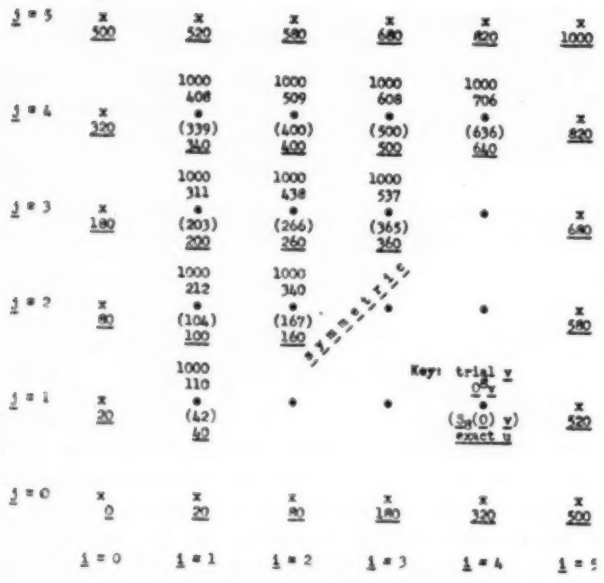


FIG. 1. Solution of Poisson's difference equation $\omega u - 20 = u$ on a 5×5 net.

cedure, we can derive a satisfactory approximation to the value of $1/T_m(1/\lambda_1)$. Since λ_1 is only slightly less than 1, write

$$1/\lambda_1 = 1+\zeta. \tag{20}$$

Now

$$\operatorname{arccosh}(1+\zeta) = (2\zeta)^{\frac{1}{2}}(1-\tfrac{1}{12}\zeta+\tfrac{3}{160}\zeta^2+\cdots).$$

If $\zeta \ll 1$, we can write [compare (18)],

$$T_m(1+\zeta) \approx \cosh m(2\zeta)^{\frac{1}{2}} = \tfrac{1}{2}e^{m\sqrt{(2\zeta)}} + \tfrac{1}{2}e^{-m\sqrt{(2\zeta)}}.$$

The last term here will be negligible whenever $1/T_m(1+\zeta)$ is small; it is negligible to 4-significant-figure accuracy whenever $1/T_m(1+\zeta)<0.02$, and all cases of interest fall within this range. Neglecting this last term, we find

$$1/T_m(1/\lambda_1) = 1/T_m(1+\zeta) \approx 2e^{-m\sqrt{(2\zeta)}}. \tag{21}$$

If we want this to be a small number ϵ, we must choose an order m of polynomial given by

$$m = (-\log\tfrac{1}{2}\epsilon)/(2\zeta)^{\frac{1}{2}}. \tag{22}$$

For a square net of $p\times p = N$ points, $\zeta \approx 1-\lambda_1 = \frac{1}{2}\pi^2/N$, and

$$m = (N/\pi^2)^{\frac{1}{2}}(-\log\tfrac{1}{2}\epsilon). \tag{23}$$

This polynomial order may be directly compared with the required power (13) of the operator O^n used in the iteration procedure. The ratio

$$\frac{n}{m} = \frac{2}{\pi}\cdot N^{\frac{1}{2}}\cdot\frac{-\log\epsilon}{-\log\frac{1}{2}\epsilon}. \tag{24}$$

Of the three factors, the first and third are of the order of unity, and the large factor is $N^{\frac{1}{2}}$. The ratio (24) is essentially the ratio of the amount of work involved in the two methods, since building up the polynomial $S_m(O)$ is less labor than any single application of O. Hence we conclude that *the use of the Tschebyscheff operator is capable of cutting machine time by a large factor of the order of $N^{\frac{1}{2}}$, where N is the total number of points in the net.* For a three-dimensional cube, the formula corresponding to (24) contains $N^{\frac{1}{3}}$ in place of $N^{\frac{1}{2}}$.

EXAMPLE

As a short example that will illustrate this method, let us consider Poisson's equation on the 5×5 network of Fig. 1, in which the boundary points are indicated by crosses, the internal points by dots. With $h=1$, the underlined values, given by

$$u = 20(i^2+j^2),$$

are an exact solution of both the differential equation

$$\Delta u - 80 = 0, \tag{25}$$

and the corresponding difference equation [compare Eq. (2)]

$$\omega u_{ij} - 20 = u_{ij}. \tag{26}$$

Let us now solve the difference equation (26) numerically by using the boundary values given in Fig. 1 but starting with a grossly erroneous trial function v having the value 1000 at all interior points. The operator O is [compare Eq. (4)]

$$Ov = \omega v - 20. \tag{27}$$

In this case the 16 eigenvectors of ω are given by

$$\sin(ki\pi/5)\sin(lj\pi/5), \quad (k,l=1,2,3,4) \tag{28}$$

with eigenvalues

$$\begin{aligned} \lambda_1, \lambda_{16} &= \pm 0.809, \\ \lambda_2, \lambda_3, \lambda_{14}, \lambda_{15} &= \pm 0.559, \\ \lambda_4, \lambda_{13} &= \pm 0.309, \\ \lambda_5, \lambda_6, \lambda_{11}, \lambda_{12} &= \pm 0.250, \\ \lambda_7, \lambda_8, \lambda_9, \lambda_{10} &= 0. \end{aligned} \tag{29}$$

The error $\phi = v-u$ [compare Eq. (5)] in the trial function v, when expanded as in (9) in terms of the eigenvectors (28), has as its leading term

$$c_1\phi^{(1)} = 1076\sin(i\pi/5)\sin(j\pi/5). \tag{30}$$

This term is the troublesome one to remove by numeri-

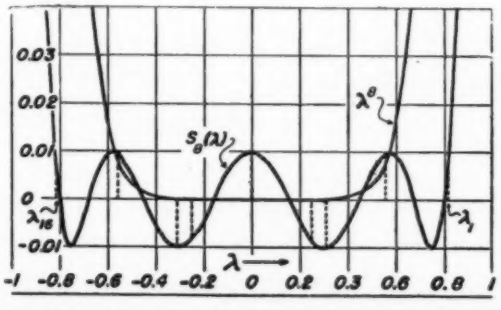


FIG. 2. Plot showing λ^8, $S_8(\lambda)$ and the eigenvalues (broken lines) for the net of Fig. 1.

cal procedures. The coefficient of $\phi^{(16)}$, on the other hand, happens to be zero, so this eigenvector will not prove troublesome.

If we apply the operator O eight times to v, we get the function O^8v shown in the second line of Fig. 1. (The computations were actually made by carrying two more figures to avoid rounding-off errors.) Figure 2 shows a plot of λ^8 with the eigenvalues (29) indicated by the broken lines. This plot shows that the coefficients of all eigenvectors in ϕ will be reduced in O^8v to less than one percent of their initial values, *except* c_1, which will be multiplied by the factor $\lambda_1^8=(0.809)^8=0.183$. The maximum error in O^8v in Fig. 1 is 180, which is consistent with (30) multiplied by λ_1^8. To achieve a value $\lambda_1^n=0.01$ would require $n=22$, i.e., 22 iterations.

Figure 3 is a plot along the diagonal $i=j$ of the trial function v and the results of the eight iterations, showing the manner of convergence to the true solution u.

Now let us apply the Tschebyscheff scheme. We see from Table I of reference 2 that $1/T_8(1/0.818)=1/T_8(1/9/11)=0.0106$; hence the operator [compare Eq. (17)]

$$S_8(O)=T_8(1\tfrac{1}{9}O)/T_8(1\tfrac{1}{9}) \tag{31}$$

should reduce all coefficients to one percent or less from $\lambda=-0.818$ to $+0.818$. Since this region includes λ_1 (compare Fig. 2), the operator (31) should reduce the coefficient in (30) to less than one percent.

We easily compute that

$$S_8(O)v=6.7620\,O^8v-9.0532\,O^6v+3.7878\,O^4v \\ -0.5071\,O^2v+0.0106\,v.$$

Formation of this linear combination gives the values in parentheses in Fig. 1. The maximum error in the function $S_8(O)v$ is 7, so that the expected reduction of (30) to less than one percent has been achieved. The difference between $S_8(O)v$ and u is too small to show on the plot of Fig. 3.

COMMENTS

The techniques for machine application of the Tschebyscheff polynomial suggested in reference 2, including the factoring of the polynomial into polynomials of order six or eight, are applicable also to boundary-value problems.

In order to cut the order of polynomial required [compare Eq. (22)], it may be desirable not to try to remove $\phi^{(1)}$, or even $\phi^{(1)}$ and $\phi^{(2)}$, from the error, but to concentrate on removing $\phi^{(3)}, \cdots, \phi^{(N)}$. One can do this by using an appropriate shift of scale and origin of the Tschebyscheff polynomial as described in reference 2. The following process similar to the "harmonic analysis" of reference 2, can then be used to remove $\phi^{(1)}$ and $\phi^{(2)}$.

Suppose we have arrived at the function

$$v=u+a_1\phi^{(1)}+a_2\phi^{(2)}. \tag{32a}$$

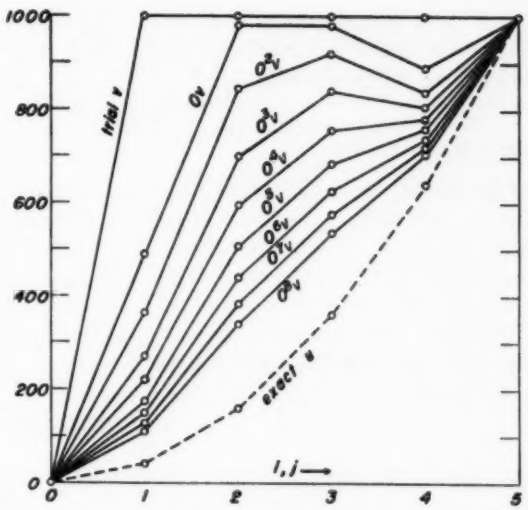


FIG. 3. Plot along the diagonal of symmetry, $i=j$, of Fig. 1.

Then [compare Eqs. (6) and (10)], we can compute

$$Ov=u+\lambda_1a_1\phi^{(1)}+\lambda_2a_2\phi^{(2)}, \tag{32b}$$

$$v-Ov=a_1(1-\lambda_1)\phi^{(1)}+a_2(1-\lambda_2)\phi^{(2)}. \tag{33a}$$

Write this last equation for the moment as

$$w=c_1\phi^{(1)}+c_2\phi^{(2)}. \tag{33b}$$

This function w is of the form of Eq. (12a) of reference 2. We can now operate on w twice by ω and follow the procedure on p. 1330 of reference 2 to determine λ_1, λ_2, $c_1\phi^{(1)}$, and $c_2\phi^{(2)}$. Since $a_1(1-\lambda_1)=c_1$ and $a_2(1-\lambda_2)=c_2$, we can eliminate the error from (32) and determine u.

The procedure just mentioned can also be used to obtain estimates of λ_1 and λ_2 from a coarse net, as discussed in reference 2, for guidance in choosing the order of polynomial required on a refined net.

The device used in reference 3 of defining the operator ω in O so that, in traversing the net with the improvement formula, the improved values Ov are used whenever they are available, doubles the value of $1-\lambda_1$ and hence, according to (12), cuts the number of necessary iterations by a factor of two in the iteration procedure. The same device can be used in the Tschebyscheff method, in which case, according to (22), it cuts the order of polynomial required by $\sqrt{2}$. However in machine work (as is the case with IBM machines) this use of improved values may not be very convenient.

We might briefly discuss one other type of example. In the case of axisymmetrical boundary-value problems,[4] the equation

$$\frac{\partial^2u}{\partial z^2}+\frac{\partial^2u}{\partial\rho^2}+\frac{K}{\rho}\frac{\partial u}{\partial\rho}=0 \tag{34}$$

arises with $K=1$, 3, and 5 in various cases. Here ρ is the radial coordinate and z the coordinate parallel to

[4] Shortley, Weller, Darby, and Gamble, J. Appl. Phys. 18, 116 (1947).

the axis. If a square net is laid down with points at $\rho=ih$, $z=jh$, the differential equation (34) is approximated by the difference equation

$$\Omega u = u,$$

where Ω is defined by

$$\Omega v_{ij}=\tfrac{1}{4}v_{i,j+1}+\tfrac{1}{4}v_{i,j-1}+[\tfrac{1}{4}+(K/8i)]v_{i+1,j} +[\tfrac{1}{4}-(K/8i)]v_{i-1,j}.$$

This operator (with certain modifications near the axis[4]) can be used as an improvement operator like O above. It can be represented by a real matrix that satisfies the conditions (see footnote 3 of reference 2) for having real eigenvalues. Hence the employment of a Tschebyscheff polynomial in Ω can speed up the convergence in the same way as for Poisson's equation.

JOURNAL OF APPLIED PHYSICS VOLUME 24, NUMBER 4 APRIL, 1953

Optimum Nonlinear Filters

L. A. ZADEH
Department of Electrical Engineering, Columbia University, New York, New York
(Received September 11, 1952)

The theory of optimum nonlinear filters outlined in this paper is based on the consideration of a sequence of classes of nonlinear filters, designated as $\mathfrak{N}_1$, $\mathfrak{N}_2$, $\mathfrak{N}_3,\cdots$, such that each class in the sequence includes all the preceding classes and, furthermore, the class of linear filters is a subclass of every class in the sequence. A filter of class $\mathfrak{N}_m$ is described in terms of a characteristic function which involves m age variables and m values of the input time-function. The input-output relationship for a filter of class $\mathfrak{N}_m$ has the form of an m-fold integral of the characteristic function with respect to the m age variables. It is shown that the characteristic function of the optimum filter (in the least squares sense) within the class $\mathfrak{N}_m$ satisfies a linear integral equation of $2m$th order. The optimization of filters of class $\mathfrak{N}_1$ is treated in detail, and methods of approximate realization of such filters in the form of nonlinear delay line filters and power series filters are indicated. The results are extended to the case of nonstationary time series.

1. INTRODUCTION

THE theory of optimum linear filters has been the subject of numerous investigations since the publication of Wiener's classic monograph[1] on the prediction and smoothing of stationary time series. By contrast, relatively little work has been done on the optimization of nonlinear filters, although it has long been recognized that, in principle, better results can be achieved with filters of this type than with linear filters. An important deterrent to the study of nonlinear filters has been the fact that, in general, nonlinear systems are not susceptible of a strictly analytical treatment and therefore do not provide the system theorist with a fruitful field for purely theoretical investigation. In recent years, however, the advent of large scale digital computers and other mechanized means of computation has profoundly influenced the basic philosophy of system design and analysis. Thus, in virtue of the availability of machine computers, it has become sufficient to carry the analytical treatment only to a point where the problem is reduced to mathematical operations which can be handled by such computers. This development has made it practicable to study nonlinear systems which cannot be completely analyzed by purely analytical means.

The use of machine computers is implicit in the important contributions by Singleton[2] and White[3] to the theory of optimum nonlinear filters. Essentially, Singleton and White consider the class of nonlinear filters which are characterized by input-output relationships of the form

$$v(t)=f[u(t), u(t-T),\cdots, u(t-(n-1)T)], \qquad (1)$$

where $u(t)$ denotes the input, $v(t)$ is the output, T is a constant, and f is an arbitrary real function, and determine the function f in such a way as to minimize the mean square error (or the probability of error). In this method, the determination of the minimizing function f requires, in general, the knowledge of $2n$th order joint probability distribution function for the signal and noise components of the input.

The purpose of this paper is to outline a more flexible approach which is based on the consideration of a certain system of classes of nonlinear filters.[4] Specifically, we consider a sequence of classes, designated as $\mathfrak{N}_1$, $\mathfrak{N}_2$, $\mathfrak{N}_3,\cdots$, such that each class in the sequence includes all the preceding classes and, furthermore, each

[1] N. Wiener, "The extrapolation, interpolation and smoothing of stationary time series," Report to the Services 19, Research Project DIC-6037, M.I.T. (February, 1942). Published in book form by John Wiley and Sons, Inc., New York, 1949.

[2] H. E. Singleton, "Theory of nonlinear transducers," Tech. Report No. 160, M.I.T. Research Laboratory of Electronics (1950).

[3] W. D. White, Proc. Inst. Radio Engrs. 39, 303 (1951).

[4] This and other systems of classes of nonlinear two-poles are discussed in greater detail in a paper by the writer, "A contribution to the theory of nonlinear systems," to be published in J. Franklin Inst.

class includes the class of linear filters. In virtue of these inclusion properties, the optimum filter (in any specified sense and for any specified purpose) within a class $\mathfrak{N}_m$ is, in general, superior to, and more complex than, the optimum filters within the classes $\mathfrak{N}_{m-1}$, $\mathfrak{N}_{m-2}, \cdots$, $\mathfrak{N}_0$ and, *a fortiori*, the class of linear filters. Furthermore, the amount of statistical data needed for determining the optimum filter within a class $\mathfrak{N}_m$ increases with m. Specifically, if the optimum filter is defined as one which minimizes the mean square error, it is found that the determination of the optimum filter within a class $\mathfrak{N}_m$ requires the knowledge of $2m$th order probability distribution functions for the signal and noise.

In employing the approach described in the sequel, the designer of an optimum nonlinear filter would select a class $\mathfrak{N}_m$, $m=1, 2, \cdots$, (or possibly a subclass of $\mathfrak{N}_m$) consistent with the allowable complexity of the filter and the available statistical data about the signal and noise, and then would proceed to determine the optimum filter within the class $\mathfrak{N}_m$. Once the characteristic function [see Eq. (2) *et seq.*] of the optimum filter has been determined, the filter can be realized in one of a number of canonical forms for filters of class $\mathfrak{N}_m$. One such form consists of an m-dimensional delay-line filter with zero-memory nonlinear $(m+1)$-poles serving as weighting functions.

The classes $\mathfrak{N}_1$, $\mathfrak{N}_2$, $\mathfrak{N}_3, \cdots$ are defined as the collections of all filters whose input-output relationships can be expressed in the following forms:

$$\mathfrak{N}_1: v(t)=\int_0^\infty K[u(t-\tau), \tau]d\tau \tag{2}$$

$$\mathfrak{N}_2: v(t)=\int_0^\infty\int_0^\infty K[u(t-\tau_1), \times u(t-\tau_2), \tau_1, \tau_2]d\tau_1 d\tau_2 \tag{3}$$

$$\mathfrak{N}_3: v(t)=\int_0^\infty\int_0^\infty\int_0^\infty K[u(t-\tau_1), \times u(t-\tau_2), u(t-\tau_3), \tau_1, \tau_2, \tau_3]d\tau_1 d\tau_2 d\tau_3, \tag{4}$$

. . . . . .

where $u(t)$ is the input to the filter, $v(t)$ is the output, and the τ's are the variables of integration. The function

$$K[u(t-\tau_1), u(t-\tau_2), \cdots, u(t-\tau_m), \tau_1, \tau_2, \cdots, \tau_m],$$

$m=1, 2, 3, \cdots$, is any real function of its arguments (possibly including delta-functions of various orders) and is called the *characteristic function* of the filter with which it is associated.

The theory outlined in this paper does not lead to explicit expressions for the characteristic functions of optimum filters within the classes $\mathfrak{N}_1$, $\mathfrak{N}_2$, $\mathfrak{N}_3, \cdots$. Rather, it leads to a sequence of linear integral equations which are satisfied by the respective characteristic functions. Thus, the determination of the optimum filter within a specified class $\mathfrak{N}_m$ is reduced to the solution of a linear integral equation of the first kind and $2m$th order. In general, the solution of this equation and hence the determination of the optimum filter requires the use of a machine computer.

In what follows, in order to avoid a treatment of excessive length we shall confine our attention to the optimization of filters of class $\mathfrak{N}_1$. This will not entail a significant loss in generality since the optimization of filters of higher classes closely parallels that of filters of class $\mathfrak{N}_1$. In fact, once the integral equation satisfied by the characteristic function of the optimum filter within the class $\mathfrak{N}_1$ has been derived, it is a trivially simple matter to generalize it to filters of class $\mathfrak{N}_m (m>1)$.

2. PROPERTIES OF FILTERS OF CLASS $\mathfrak{N}_1$

In this section, we shall establish certain basic properties of filters of class $\mathfrak{N}_1$ that are pertinent to the optimization and realization of such filters. It will be assumed throughout that the input to the filter is applied at rest and that the output is zero in the absence of the input.

First, we note that the class of linear filters is a subclass of $\mathfrak{N}_1$. This follows at once from the fact that the input-output relationship for any linear filter can be expressed in the form

$$v(t)=\int_0^\infty W(\tau)u(t-\tau)d\tau, \tag{5}$$

where $W(t)$ is the impulsive response of the filter. Clearly, this relation is a special case of the input-output relationship

$$v(t)=\int_0^\infty K[u(t-\tau), \tau]d\tau \tag{6}$$

which characterizes the filters of class $\mathfrak{N}_1$. Consequently, the class of linear filters is a subclass of $\mathfrak{N}_1$, with the characteristic function of filters within this subclass having the form

$$K[u(t-\tau), \tau]=W(\tau)u(t-\tau). \tag{7}$$

Next, we note that any zero-memory (i.e., noninertial) nonlinear filter is of class $\mathfrak{N}_1$. The input-output relationship for this type of filter reads

$$v(t)=f[u(t)], \tag{8}$$

where f is a real function. On writing this relation in the form

$$v(t)=\int_0^\infty \delta(\tau)f[u(t-\tau)]d\tau, \tag{9}$$

where $\delta(\tau)$ is the unit impulse function (delta-function), it is seen that (9) is a special case of (6) and hence the class of zero-memory filters is a subclass of $\mathfrak{N}_1$.

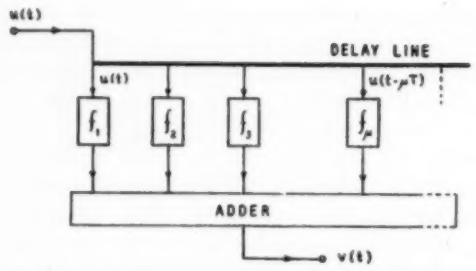


FIG. 1. A nonlinear delay-line filter of class $\mathfrak{N}_1$.

An important property of filters of class $\mathfrak{N}_1$ concerns their characterizability in terms of responses to step functions. Thus, let the input to a filter of class $\mathfrak{N}_1$ be a step function of amplitude α, i.e.,

$$u(t)=\alpha 1(t), \tag{10}$$

where $1(t)$ denotes the unit step function. On substituting (10) in (6), the expression for the response to the step function is found to be

$$v(t)=\int_0^t K[\alpha,\tau]d\tau, \tag{11}$$

which upon differentiation with respect to t yields

$$\frac{dv(t)}{dt}=K[\alpha,t]. \tag{12}$$

This relation states that:

$K[\alpha,t]$ = derivative of the response of the filter to a step function of amplitude α. (13)

Consequently, a filter of class $\mathfrak{N}_1$ is completely characterized by its responses to a set of step functions $\{\alpha 1(t)\}$ whose amplitudes $\{\alpha\}$ range over all real values.

It is of interest to observe that this property of filters of class $\mathfrak{N}_1$ is a generalization of a well-known property of linear filters, which is that a linear filter is completely characterized by its response to a single step function. Furthermore, the impulsive response of a linear filter may be expressed as

$W(t)$ = derivative of the response of the filter to the unit step function, (14)

which is seen to be a special case of (13).

There are two subclasses of $\mathfrak{N}_1$ that are worthy of special consideration because they furnish a convenient means for the approximate realization of any filter of class $\mathfrak{N}_1$. The first of these is the subclass of nonlinear delay-line filters of the form shown in Fig. 1. The input-output relationship for this type of filter[5] is of the form

$$v(t)=\sum_{\mu=0}^{m} f_\mu[u(t-\mu T)], \tag{15}$$

where the f_μ are arbitrary real functions, T is a positive constant, and m is a positive integer, not necessarily finite. The characteristic function in this case reads

$$K[u(t-\tau),\tau]=\sum_{\mu=0}^{m}\delta(\tau-\mu T)f_\mu[u(t-\tau)]. \tag{16}$$

The block labeled f_μ in Fig. 1 represents a zero-memory nonlinear two-pole characterized by the input-output relationship

$$v(t)=f_\mu[u(t)]. \tag{17}$$

The input to this two-pole is $u(t-\mu T)$, where $u(t)$ is the input to the filter.

The subclass of nonlinear delay-line filters is dense in the class $\mathfrak{N}_1$, in the sense that any filter of class $\mathfrak{N}_1$ is either of this type or can be approximated as closely as desired by a filter of this type. This implies that any filter of class $\mathfrak{N}_1$ can be realized approximately in the form shown in Fig. 1.

Another important subclass of $\mathfrak{N}_1$ is made up of filters whose input-output relationship is of the form

$$v(t)=\sum_{\mu=1}^{m}\int_0^\infty W_\mu(\tau)[u(t-\tau)]^\mu d\tau, \tag{18}$$

where the coefficients $W_\mu(\tau)$ are arbitrary real functions of τ. For convenience, a filter of this type will be referred to as *power series* filter of order m.

The power series type of filter is readily realized in the form shown in Fig. 2. The filter consists of m channels, a typical channel being made up of a nonlinear device which yields a power of $u(t)$, say $[u(t)]^\mu$, and a linear filter whose impulsive response is the coefficient of $[u(t-\tau)]^\mu$ in (18), i.e., $W_\mu(t)$. It will be noted that the linear filter is a special case of the power series filter, resulting when $W_2(t)=W_3(t)=\cdots=W_m(t)\equiv 0$.

The characteristic function of a power series filter is of the form

$$K[u(t-\tau),\tau]=\sum_{\mu=1}^{m}W_\mu(\tau)[u(t-\tau)]^\mu. \tag{19}$$

By approximating the characteristic function of a specified filter of class $\mathfrak{N}_1$ in this form, one obtains a power series filter which approximates the specified filter. The error in approximation can be made as small as desired, which implies that the subclass of power series filters is dense in the class $\mathfrak{N}_1$.

In addition to the subclasses discussed above, there are a number of others which might be of use in special applications. These subclasses are associated with various special forms of the characteristic function.

3. OPTIMIZATION OF FILTERS OF CLASS $\mathfrak{N}_1$

In general, the input $u(t)$ is a mixture, not necessarily additive, of a signal $s(t)$ and a noise $n(t)$. For simplicity, we shall assume that $s(t)$ and $n(t)$ are ergodic random functions. This restriction will be removed in Sec. 6.

[5] Some results pertaining to the stability of a related (pulsed) type of filter were recently reported by R. L. Sternberg, Bull. Am. Math. Soc. 58, 389 (1952), abstract only.

To assess the performance of a filter, we shall use the conventional mean-square-error criterion. With this criterion, the optimum filter within $\mathfrak{N}_1$ is that filter of class $\mathfrak{N}_1$ which minimizes the mean square value of the difference between the desired output and the actual output of the filter. The desired output is usually the signal $s(t)$, delayed or advanced by β-seconds. In some instances, the desired output is a derivative of $s(t)$ or, more generally, some functional transform of $s(t)$. It will be convenient to represent the desired output as the result of operating with a filter of class $\mathfrak{N}_1$ on the signal $s(t)$. Thus, on denoting the desired output by $v^*(t)$ and indicating the characteristic function of the filter which yields it by $K^*[s(t-\tau), \tau]$, we have,

$$v^*(t)=\int_{-\infty}^{\infty} K^*[s(t-\tau), \tau]d\tau. \tag{20}$$

In terms of this representation, if the desired output is $s(t+\beta)$, for example, then

$$K^*[s(t-\tau), \tau]=\delta(\tau+\beta)s(t-\tau); \tag{21}$$

and if

$$v^*(t)=s^2(t) \tag{22}$$

then

$$K^*[s(t-\tau), \tau]=\delta(\tau)[s(t-\tau)]^2. \tag{23}$$

Note that the lower limit in (20) is $-\infty$ rather than 0 as in (6), since the operator defined by (20) need not be physically realizable.

The difference between the actual output, $v(t)$, and the desired output, $v^*(t)$, is the error $\epsilon(t)$,

$$\epsilon(t)=v(t)-v^*(t). \tag{24}$$

The performance of the filter is assessed in terms of the mean square value of the error, σ^2,

$$\sigma^2=\langle \epsilon^2(t)\rangle_{Av}. \tag{25}$$

In this expression it is immaterial whether $\langle\ \rangle_{Av}$ is interpreted as the ensemble or time average, since $s(t)$ and $n(t)$ are assumed to be ergodic.

On substituting (6) and (20) in (24), squaring the resulting expression and averaging, we obtain

$$\begin{aligned}\sigma^2=&\int_0^{\infty}\int_0^{\infty}\langle K[u(t-\tau_1), \tau_1]\\ &\times K[u(t-\tau_2), \tau_2]\rangle_{Av}d\tau_1 d\tau_2\\ &-2\int_{-\infty}^{\infty}\int_0^{\infty}\langle K[u(t-\tau_1), \tau_1]\\ &\times K^*[s(t-\tau_2), \tau_2]\rangle_{Av}d\tau_1 d\tau_2\\ &+\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\langle K^*[s(t-\tau_1), \tau_1]\\ &\times K^*[s(t-\tau_2), \tau_2]\rangle_{Av}d\tau_1 d\tau_2,\end{aligned} \tag{26}$$

where τ_1 and τ_2 are the variables of integration. The third term in (26) does not involve K and hence is irrelevant to the determination of the minimizing characteristic function. The quantity that need be minimized, then, consists of the first two terms in (26). This quantity will be denoted by I.

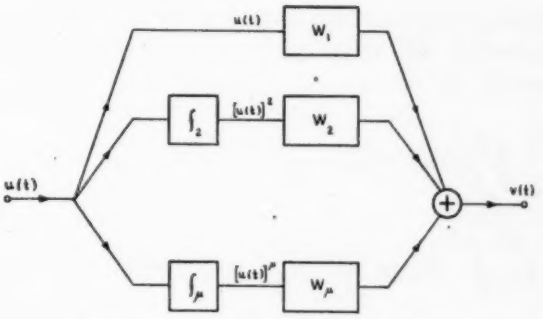


FIG. 2. A power series filter of class $\mathfrak{N}_1$.

To evaluate the averages involved in (26), it is necessary to know the joint probability density functions for $u(t-\tau_1)$ and $u(t-\tau_2)$, and $u(t-\tau_1)$ and $s(t-\tau_2)$. These are denoted by $p(u_1, u_2; \tau_1-\tau_2)$ and $p(u_1, s_2; \tau_1-\tau_2)$, respectively, where u_1, u_2, and s_2 are abbreviations for $u(t-\tau_1)$, $u(t-\tau_2)$, and $s(t-\tau_2)$. (For simplicity, the same symbol p is associated with different probability density functions.) It will be noted that the probability densities in question are functions of $\tau_1-\tau_2$, rather than τ_1 and τ_2, in virtue of the ergodicity of $s(t)$ and $n(t)$. It should also be noted that $p(u_1, u_2; \tau_1-\tau_2)$ is a symmetric function in the variables u_1, τ_1 and u_2, τ_2.[6] [This property is used in (28).]

In terms of these probability distributions, the expression for the quantity to be minimized, I, reads

$$\begin{aligned}I=&\int_0^{\infty}\int_{-\infty}^{\infty}\int_0^{\infty}\int_{-\infty}^{\infty}K(u_1, \tau_1)K(u_2, \tau_2)\\ &\times p(u_1, u_2; \tau_1-\tau_2)du_1 d\tau_1 du_2 d\tau_2\\ &-2\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\int_0^{\infty}\int_{-\infty}^{\infty}K(u_1, \tau_1)K^*(s_2, \tau_2)\\ &\times p(u_1, s_2; \tau_1-\tau_2)du_1 d\tau_1 ds_2 d\tau_2.\end{aligned} \tag{27}$$

The minimization of I is readily achieved by calculating the variation in I, δI, corresponding to an admissible variation $\epsilon\delta K$ in K, and setting the coefficient of ϵ in δI equal to zero. Straightforward calculation yields

$$\begin{aligned}\delta I=&2\epsilon\int_0^{\infty}\int_{-\infty}^{\infty}\delta K(u_1, \tau_1)du_1 d\tau_1\\ &\times\left\{\int_0^{\infty}\int_{-\infty}^{\infty}K(u_2, \tau_2)p(u_1, u_2; \tau_1-\tau_2)du_2 d\tau_2\right.\\ &\left.-\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}K^*(s_2, \tau_2)p(u_1, s_2; \tau_1-\tau_2)ds_2 d\tau_2\right\}\\ &+\epsilon^2\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\int_0^{\infty}\int_{-\infty}^{\infty}\delta K(u_1, \tau_1)\delta K(u_2, \tau_2)\\ &\times p(u_1, u_2; \tau_1-\tau_2)du_1 d\tau_1 du_2 d\tau_2.\end{aligned} \tag{28}$$

[6] James, Nichols, and Phillips, *Theory of Servomechanisms* (McGraw-Hill Book Company, Inc., New York, 1947), Chap. 6.

On requiring that the coefficient of ϵ be zero for all admissible $\epsilon\delta K(u_1, \tau_1)$, we obtain the desired integral equation for K:

$$\int_0^\infty \int_{-\infty}^\infty K(u_2, \tau_2) p(u_1, u_2; \tau_1-\tau_2) du_2 d\tau_2$$

$$= \int_{-\infty}^\infty \int_{-\infty}^\infty K^*(s_2, \tau_2) p(u_1, s_2; \tau_1-\tau_2) ds_2 d\tau_2, \ \tau_1 \geqslant 0 \quad (29)$$

where, to recapitulate:

$K(u, \tau) = K[u(t-\tau), \tau]$
= characteristic function of the optimum filter

$K^*(s, \tau) = K[s(t-\tau), \tau]$ = characteristic function of the filter yielding the desired output,

$u(t)$ = input to filter,

$s(t)$ = signal component of $u(t)$,

$u_1 = u(t-\tau_1), \quad u_2 = u(t-\tau_2), \quad s_2 = s(t-\tau_2)$,

$p(u_1, u_2; \tau_1-\tau_2)$ = joint probability density for u_1 and u_2,

$p(u_1, s_2; \tau_1-\tau_2)$ = joint probability density for u_1 and s_2.

It remains to be shown that the solution of (29) actually minimizes I. This follows at once from the fact that for the characteristic function satisfying (29), the expression for δI [Eq. (28)] reduces to the second term in (28), which is a positive quantity since it is equal to

$$\left\langle \left[\int_0^\infty \delta K[u(t-\tau), \tau] d\tau \right]^2 \right\rangle_{Av}.$$

Consequently, any characteristic function different from that satisfying (29) will result in a value of I greater than that corresponding to the solution of (29). Thus, the solution of (29) gives a true minimum of I.

To summarize, the characteristic function of the optimum filter of class $\mathfrak{N}_1$ is the solution of the linear integral equation (29). The statistical data necessary for the determination of the optimum filter consist of the second-order probability distribution function for the input, and the joint probability distribution for the input and its signal component.

It will be noted that Eq. (29) is of the Wiener-Hopf type in the variables τ_1 and τ_2. It is theoretically possible, though rarely practicable, to reduce its order by one through the use of the factorization technique. A more practical, though only approximate, approach which results in a reduction of (29) to an integral equation of the Fredholm type, is to apply the Fourier transformation to both sides of (29), disregarding the restriction that (29) holds only for positive τ_1. [The disregarding of this restriction is in effect equivalent to allowing the filter to have an infinite time-delay.] This leads to the following relation between the Fourier transforms of the quantities involved in (29):

$$\int_{-\infty}^\infty K(u_2, \omega) p(u_1, u_2; \omega) du_2 = \int_{-\infty}^\infty K^*(s_2, \omega) p(u_1, s_2; \omega) ds_2, \quad (30)$$

where $K(u_2, \omega)$ is the Fourier transform of $K(u_2, \tau_2)$, $p(u_1, u_2; \omega)$ is the Fourier transform of $p(u_1, u_2; \tau_1-\tau_2)$, etc. For simplicity, the same symbols are used for the functions and their respective Fourier transforms.

Equation (30) is of the Fredholm type and of the first kind. In order to solve this equation for $K(u_2, \omega)$, it is necessary, in general, to employ machine computation. Once $K(u_2, \omega)$ has been computed, the characteristic function $K(u_2, \tau_2)$ may be obtained by applying the inverse Fourier transformation to $K(u_2, \omega)$. In general, the $K(u_2, \tau_2)$ function so obtained would not vanish for negative τ_2, as it should in the case of a physically realizable filter. However, it is possible, in general, to realize approximately $K[u(t-\tau), \tau-T_d]$, where T_d is a sufficient large positive constant, by neglecting the part of $K[u(t-\tau), \tau-T_d]$ which extends over negative τ. In this way, the characteristic function yielded by (30) can be approximately realized to within a constant time-delay T_d.

The approach sketched above is suited for use in conjunction with automatic integral equation solvers and Fourier transformers. In those cases where a large scale digital computer is available, the reduction of (29) to (30) would not, in general, be worth while, since it would not be difficult to program the solution of (29) on such a computer.

For purposes of illustration, we shall consider a simple filtering problem in the light of the foregoing analysis. The characteristics of the signal and noise are taken to be as follows.

Noise, $n(t)$: Gaussian, with correlation function $\psi_n(\tau) = e^{-\alpha|\tau|}$, where α is a positive constant. The second probability distribution function is given by

$$p_n(n_1, n_2) = \frac{1}{2\pi[1-\exp(-2\alpha|\tau_1-\tau_2|)]^{\frac{1}{2}}} \times \exp\left\{ -\frac{n_1^2 - 2n_1 n_2 \exp(-\alpha|\tau_1-\tau_2|) + n_2^2}{2[1-\exp(-2\alpha|\tau_1-\tau_2|)]} \right\} \quad (31)$$

where $n_1 = n(t-\tau_1)$ and $n_2 = n(t-\tau_2)$.

Signal, $s(t)$: Random square wave, independent of $n(t)$, with $s(t)$ taking the values 1 and 0 with equal probability over intervals of random length l, the probability density of l being $\beta e^{-\beta l}$ (Poisson process), where $1/\beta = \langle l \rangle_{Av}$. It can readily be shown[5] that the probability that $t-\tau_1$ and $t-\tau_2$ lie within the same

interval is

$$Q(\tau_1-\tau_2)=\exp(-\beta|\tau_1-\tau_2|). \quad (32)$$

Consequently, the second probability distribution function is given by

$$\begin{aligned}p_s(s_1, s_2)&=\tfrac{1}{4}[1-e^{-\beta|\tau_1-\tau_2|}]\\&\times[\delta(s_1)\delta(s_2-1)+\delta(s_1-1)\delta(s_2)]\\&+\tfrac{1}{4}[1+e^{-\beta|\tau_1-\tau_2|}]\\&\times[\delta(s_1)\delta(s_2)+\delta(s_1-1)\delta(s_2-1)], \quad (33)\end{aligned}$$

where $s_1=s(t-\tau_1)$, $s_2=s(t-\tau_2)$, and $\delta(s)$ is the delta-function.

Input, $u(t)$: $u(t)=s(t)+n(t)$.

Desired output, $v^*(t)$: $v^*(t)=s(t)$. Consequently, $K^*(s_2, \tau_2)=\delta(\tau_2)s_2$ and the right-hand member of (29) reduces to

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}K^*(s_2, \tau_2)p(u_1, s_2; \tau_1-\tau_2)ds_2d\tau_2 = \int_{-\infty}^{\infty}s_2p(u_1, s_2; \tau_1)ds_2. \quad (34)$$

The kernel of the integral equation (29) is related to the second probability distribution functions for the signal and noise by the two-dimensional convolution integral

$$p(u_1, u_2; \tau_1-\tau_2)=\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}p_n(u_1-s_1, u_2-s_2)\,p_s(s_1, s_2)ds_1ds_2. \quad (35)$$

On substituting (31) and (33) in (35) and performing the integration, we find

$$\begin{aligned}p(u_1, u_2; \tau_1-\tau_2)=&\frac{1-\exp(-\beta|\tau_1-\tau_2|)}{8\pi[1-\exp(-2\alpha|\tau_1-\tau_2|)]^{\frac{1}{2}}}\left\{\exp\left[-\frac{u_1^2-2u_1(u_2-1)e^{-\alpha|\tau_1-\tau_2|}+(u_2-1)^2}{2(1-e^{-2\alpha|\tau_1-\tau_2|})}\right]\right.\\&\left.+\exp\left[-\frac{(u_1-1)^2-2(u_1-1)u_2e^{-\alpha|\tau_1-\tau_2|}+u_2^2}{2(1-e^{-2\alpha|\tau_1-\tau_2|})}\right]\right\}\\&+\frac{1+\exp(-\beta|\tau_1-\tau_2|)}{8\pi[1-\exp(-2\alpha|\tau_1-\tau_2|)]^{\frac{1}{2}}}\left\{\exp\left[-\frac{u_1^2-2u_1u_2e^{-\alpha|\tau_1-\tau_2|}+u_2^2}{2(1-e^{-2\alpha|\tau_1-\tau_2|})}\right]\right.\\&\left.+\exp\left[-\frac{(u_1-1)^2-2(u_1-1)(u_2-1)e^{-\alpha|\tau_1-\tau_2|}+(u_2-1)^2}{2(1-e^{-2\alpha|\tau_1-\tau_2|})}\right]\right\}, \quad (36)\end{aligned}$$

which is the expression for the kernel of (29).

The probability density $p(u_1, s_2; \tau_1-\tau_2)$ is given by the convolution integral

$$p(u_1, s_2; \tau_1-\tau_2)=\int_{-\infty}^{\infty}p_s(s_1, s_2)p_n(u_1-s_1)ds_1, \quad (37)$$

where $p_n(n)$ is the probability density of $n(t)$,

$$p_n(n)=1/(2\pi)^{\frac{1}{2}}\exp(-n^2/2). \quad (38)$$

Substituting (33) and (38) in (37) yields

$$\begin{aligned}&p(u_1, s_2; \tau_1-\tau_2)\\&=1/4(2\pi)^{\frac{1}{2}}\{\delta(s_2)[(1-e^{-\beta|\tau_1-\tau_2|})\exp[-\tfrac{1}{2}(u_1-1)^2]\\&\quad+(1+e^{-\beta|\tau_1-\tau_2|})\exp[-\tfrac{1}{2}u_1^2]]\\&\quad+\delta(s_2-1)[(1-e^{-\beta|\tau_1-\tau_2|})\exp[-\tfrac{1}{2}u_1^2]\\&\quad+(1+e^{-\beta|\tau_1-\tau_2|})\exp[-\tfrac{1}{2}(u_1-1)^2]]\}. \quad (39)\end{aligned}$$

On inserting this expression in (34) and performing the integration, we find:

$$\begin{aligned}&\text{right-hand member of (29)}\\&=1/4(2\pi)^{\frac{1}{2}}\{(1-e^{-\beta|\tau_1|})\exp[-\tfrac{1}{2}u_1^2]\\&\quad+(1+e^{-\beta|\tau_1|})\exp[-\tfrac{1}{2}(u_1-1)^2]\}. \quad (40)\end{aligned}$$

Thus, for the case under consideration the kernel of the integral equation (29) is given by (36), while its right-hand member is expressed by (40). It will be observed that despite there lative simplicity of the problem, the integral equation (29) is too complex for solution by analytical means.

4. INTEGRAL EQUATIONS FOR SUBCLASSES OF $\mathfrak{N}_1$

In this section, we are concerned with optimum filters within the subclasses of $\mathfrak{N}_1$ considered in Sec. 2. For the subclass of linear filters, a direct derivation of the integral equation satisfied by the characteristic function is a simple matter. It is of interest, however, to see how this equation can be deduced from the more general integral equation (29).

A. Linear Filters

The characteristic function of a linear filter is of the form $W(\tau)u(t-\tau)$, where $W(t)$ is the impulsive response of the filter. On replacing K in (29) by $W(\tau)u(t-\tau)$, multiplying both sides by u_1 and integrating,[7] we obtain

$$\int_0^{\infty}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}W(\tau_2)u_1u_2p(u_1, u_2; \tau_1-\tau_2)du_1du_2d\tau_2 = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}W^*(\tau_2)u_1s_2\times p(u_1, s_2; \tau_1-\tau_2)du_1ds_2d\tau_2, \quad (41)$$

[7] The multiplication by u_1 and subsequent integration is permissible in virtue of the fact that the admissible variation in the linear case involves only the variable τ_1.

where $W^*(t)$ is the impulsive response of the filter yielding the desired output. Now

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} u_1u_2p(u_1, u_2; \tau_1-\tau_2)du_1du_2=\psi_{uu}(\tau_1-\tau_2), \quad (42)$$

where $\psi_{uu}(\tau)$ is the autocorrelation function of the input $u(t)$; and

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} u_1s_2p(u_1, s_2; \tau_1-\tau_2)du_1ds_2=\psi_{us}(\tau_1-\tau_2), \quad (43)$$

where $\psi_{us}(\tau)$ is the cross-correlation function of the input and its signal component $s(t)$. Thus, (41) may be written as

$$\int_0^{\infty} W(\tau_2)\psi_{uu}(\tau_1-\tau_2)d\tau_2 = \int_{-\infty}^{\infty} W^*(\tau_2)\psi_{us}(\tau_1-\tau_2)d\tau_2, \quad \tau_1\geqslant 0, \quad (44)$$

which is the familiar Wiener-Hopf equation for the optimum linear filter.

B. Delay-Line Filters

In this case the characteristic function is given by (15). It is convenient to represent the desired output in terms of a characteristic function of the same form, that is, write

$$v^*(t)=\sum_{\nu=0}^{n} f_\nu^*[s(t-\nu T)]. \quad (45)$$

Then, on substituting (15) and (45) in (29) we obtain

$$\sum_{\lambda=0}^{m}\int_{-\infty}^{\infty} f_\lambda(u_2)p[u_1, u_2; (\mu-\lambda)T]du_2 = \sum_{\nu=0}^{n}\int_{-\infty}^{\infty} f_\nu^*(s_2)p[u_1, s_2; (\mu-\nu)T]ds_2, \quad \mu=0, 1, 2, \cdots, m, \quad (46)$$

which is a system of $m+1$ linear integral equations for the unknown functions $f_0, f_1, \cdots, f_m$. For the particular case of a zero-memory filter, (46) reduces to

$$f_1(u_1)p(u_1)=\int_{-\infty}^{\infty} f_1^*(s_2)p(u_1, s_2)ds_2. \quad (47)$$

C. Power Series Filters

In this case it is not permissible, as in the preceding cases, to obtain the desired integral equation by substituting the expression for the characteristic function (19) into (29). It is necessary to start with the expression for σ^2 (or rather, I), and determine the minimizing characteristic function directly.

If the desired output is written as

$$v^*(t)=\sum_{\nu=1}^{n}\int_{-\infty}^{\infty} W_\nu^*(\tau)[s(t-\tau)]^\nu d\tau, \quad (48)$$

the expression for I becomes

$$I=\sum_\mu\sum_\lambda\int_0^{\infty}\int_0^{\infty} W_\mu(\tau_1)W_\lambda(\tau_2) \times\langle[u(t-\tau_1)]^\mu[u(t-\tau_2)]^\lambda\rangle_{Av}d\tau_1d\tau_2 - 2\sum_\mu\sum_\nu\int_{-\infty}^{\infty}\int_0^{\infty} W_\mu(\tau_1)W_\nu^*(\tau_2) \times\langle[u(t-\tau_1)]^\mu[s(t-\tau_2)]^\nu\rangle_{Av}d\tau_1d\tau_2. \quad (49)$$

It is convenient to express the averages in (49) in the form of correlation functions, thus:

$$\langle[u(t-\tau_1)]^\mu[u(t-\tau_2)]^\lambda\rangle_{Av}=\psi_{uu}{}^{\mu\lambda}(\tau_1-\tau_2), \quad (50)$$

where $\psi_{uu}{}^{\mu\lambda}(\tau)$ is the cross-correlation function of $[u(t)]^\mu$ and $[u(t)]^\lambda$; and

$$\langle[u(t-\tau_1)]^\mu[s(t-\tau_2)]^\nu\rangle_{Av}=\psi_{us}{}^{\mu\nu}(\tau_1-\tau_2), \quad (51)$$

where $\psi_{us}{}^{\mu\nu}(\tau)$ is the cross-correlation function of $[u(t)]^\mu$ and $[s(t)]^\nu$.

At this point, the problem is essentially that of multiple time-series prediction[1] with linear filters, the time-series in question being $u(t)$, $[u(t)]^2, \cdots, [u(t)]^m$. Proceeding in the conventional manner, it is found that the impulsive responses $W_\lambda(\tau)$ are the solutions of the following system of integral equations

$$\sum_{\lambda=1}^{m}\int_0^{\infty} W_\lambda(\tau_2)\psi_{uu}{}^{\mu\lambda}(\tau_1-\tau_2)d\tau_2 = \sum_{\nu=1}^{n}\int_{-\infty}^{\infty} W_\nu^*(\tau_2)\psi_{us}{}^{\mu\nu}(\tau_1-\tau_2)d\tau_2, \quad \tau_1\geqslant 0, \quad \mu=1, 2, \cdots, m. \quad (52)$$

These equations may be solved by the method of undetermined coefficients.[1] It is more practical, however, to use an approximate procedure[8] similar to that used previously for the reduction of (29) to (30). This involves the application of Fourier transformation to both sides of (52) (disregarding the restriction $\tau_1\geqslant 0$), which yields the algebraic system of equations

$$\sum_{\lambda=1}^{m}H_\lambda(\omega)\phi_{uu}{}^{\mu\lambda}(\omega)=\sum_{\nu=1}^{n}H_\nu^*(\omega)\phi_{us}{}^{\mu\nu}(\omega), \quad \mu=1, 2, \cdots, m, \quad (53)$$

where $H_\lambda(\omega)$, $H_\nu^*(\omega)$, $\phi_{uu}{}^{\mu\lambda}(\omega)$, and $\phi_{us}{}^{\mu\nu}(\omega)$ denote the Fourier transforms of $W_\lambda(t)$, $W_\nu^*(t)$, $\psi_{uu}{}^{\mu\lambda}(\tau)$ and $\psi_{us}{}^{\mu\nu}(\tau)$, respectively.

[8] J. P. Costas, "Interference filtering," Tech. Report No. 185, M.I.T. Research Laboratory of Electronics (1951).

It is convenient to denote the right-hand member of (53) by $B_\mu(\omega)$, which is a known function of ω. Then, on solving the system of equations for the $H_\lambda(\omega)$, we obtain

$$H_\lambda(\omega)=\sum_{\mu=1}^{m}\theta_{uu}{}^{\mu\lambda}(\omega)B_\mu(\omega),\quad \lambda=1, 2,\cdots, m, \quad (54)$$

where the matrix $[\theta_{uu}{}^{\mu\lambda}(\omega)]$ is the inverse of the coefficient matrix $[\phi_{uu}{}^{\mu\lambda}(\omega)]$, and the $H_\lambda(\omega)$ represent the transfer functions of the linear two-poles in Fig. 2.

It will be observed that the determination of the optimum power series filter of order m requires the knowledge of $m(m+1+2n)/2$ cross-correlation functions, $\psi_{uu}{}^{\mu\lambda}$ and $\psi_{us}{}^{\mu\nu}$, $\mu=1, 2,\cdots, m$, $\nu=1, 2,\cdots, n$.

5. OPTIMUM FILTERS OF CLASS $\mathfrak{N}_m$

In this section, the integral equation (29) satisfied by the characteristic function of the optimum filter within the class $\mathfrak{N}_1$, is generalized to filters of class $\mathfrak{N}_m$. It will be recalled that filters of class $\mathfrak{N}_m$ are characterized by input-output relationship of the form

$$v(t)=\int_0^\infty\int_0^\infty\cdots\int_0^\infty K[u(t-\tau_1), u(t-\tau_2),\cdots, \times u(t-\tau_m), \tau_1, \tau_2,\cdots, \tau_m]d\tau_1 d\tau_2\cdots d\tau_m, \quad (55)$$

where K is any real function. Proceeding as in Sec. 3, one readily obtains a linear integral equation of $2m$th order which is satisfied by the characteristic function of the optimum filter within $\mathfrak{N}_m$. This equation is essentially of the same form as (29) and reads

$$\int_0^\infty\cdots\int_0^\infty\int_{-\infty}^\infty\cdots\int_{-\infty}^\infty K[u_1,\cdots, u_m, \tau_1,\cdots, \tau_m] \times p(u_1,\cdots, u_m, u_1',\cdots, u_m'; \tau_1,\cdots, \tau_m, \tau_1',\cdots, \tau_m') \times du_1\cdots du_m d\tau_1\cdots d\tau_m$$

$$=\int_{-\infty}^\infty\cdots\int_{-\infty}^\infty\int_{-\infty}^\infty\cdots\int_{-\infty}^\infty K^*[s_1,\cdots, s_m, \tau_1,\cdots, \tau_m] \times p(s_1,\cdots, s_m, u_1',\cdots, u_m'; \tau_1,\cdots, \tau_m, \tau_1',\cdots, \tau_m') \times ds_1\cdots ds_m d\tau_1\cdots d\tau_m, \quad (56)$$

where

$$s_\mu=s(t-\tau_\mu),\quad \mu=1, 2,\cdots, m$$
$$u_\mu=u(t-\tau_\mu),\quad \mu=1, 2,\cdots, m$$
$$u_\mu'=u(t-\tau_\mu'),\quad \mu=1, 2,\cdots, m$$

$K[u_1, u_2,\cdots, u_m, \tau_1, \tau_2,\cdots, \tau_m]$
= characteristic function of the optimum filter.

$K^*[s_1, s_2,\cdots, s_m, \tau_1, \tau_2,\cdots, \tau_m]$
= characteristic function of a filter of class $\mathfrak{N}_m$ which yields the desired output.

$p(u_1,\cdots, u_m, u_1',\cdots, u_m'; \tau_1,\cdots, \tau_m, \tau_1',\cdots, \tau_m')$
= joint probability density of $u_1,\cdots, u_m, u_1',\cdots, u_m'$.

$p(s_1,\cdots, s_m, u_1',\cdots, u_m'; \tau_1,\cdots, \tau_m, \tau_1',\cdots, \tau_m')$
= joint probability density of $s_1,\cdots, s_m, u_1',\cdots, u_m'$.

Needless to say, it would be impracticable to solve this equation even with the aid of machine computers when m is not a small integer. Indeed, in most practical cases it would probably be expedient to restrict the class within which the optimum filter is determined, to a subclass of $\mathfrak{N}_m$ which is dense in $\mathfrak{N}_m$. One such subclass consists of filters whose input-output relationships are of the form

$$v(t)=\sum_{\lambda,\mu,\cdots,\nu} f_{\lambda,\mu,\cdots,\nu}[u(t-\lambda T), u(t-\mu T), \times\cdots, u(t-\nu T)],\ \lambda, \mu,\cdots, \nu=0, 1,\cdots, n, \quad (57)$$

where the $f_{\lambda,\mu,\cdots,\nu}$ are arbitrary real functions and T is a constant. The integral equations satisfied by the optimal $f_{\lambda,\mu,\cdots,\nu}$ functions are related to (57) in the same manner as (46) is related to (15). A filter belonging to this subclass may be realized in the form of an m-dimensional delay-line filter, i.e., a structure consisting of m tapped delay lines (or a single delay line with multiply interconnected taps) whose taps are connected to zero-memory $(m+1)$-poles, the outputs of which represent the terms of (57). The output, $v(t)$, is obtained by adding the outputs of the zero-memory $(m+1)$-poles.

6. NONSTATIONARY TIME SERIES AND FINITE-MEMORY FILTERS

The integral equations derived in the preceding sections on the assumption of ergodicity of $s(t)$ and $n(t)$, can readily be extended to the case where either $s(t)$ or $n(t)$ or both are nonstationary. Furthermore, these equations can easily be generalized to finite-memory filters, that is, filters which operate on a finite part of the past. For simplicity of treatment, we shall, as before, confine our attention to filters of class $\mathfrak{N}_1$.

It is evident that in the case of a nonstationary input, the optimum filter should be of the time-variant type. A time-variant filter of class $\mathfrak{N}_1$ is characterized by the following form of input-output relationship:

$$v(t)=\int_0^\infty K[u(t-\tau), \tau, t]d\tau, \quad (58)$$

which differs from (6) in only that the characteristic function is dependent on t. Thus, to obtain the integral equation satisfied by the characteristic function of the optimum time-variant filter, it is sufficient to replace the time-invariant characteristic functions in (29) by their time-variant counterparts. The resulting integral equation reads

$$\int_0^\infty\int_{-\infty}^\infty K[u_2, \tau_2, t]p(u_1, u_2; \tau_1, \tau_2, t)du_2 d\tau_2 = \int_{-\infty}^\infty\int_{-\infty}^\infty K^*[s_2, \tau_2, t]p(u_1, s_2; \tau_1, \tau_2, t)ds_2 d\tau_2, \times\tau_1\gtrless 0, \quad (59)$$

where $p(u_1, u_2; \tau_1, \tau_2, t)$ and $p(u_1, s_2; \tau_1, \tau_2, t)$ are the joint probability densities for u_1 and u_2, and u_1 and s_2, respectively, and $K^*[s, \tau, t]$ is the characteristic function of the filter yielding the desired output.

For the linear subclass of $\mathfrak{N}_1$, (59) reduces to the integral equation[9]

$$\int_0^\infty W(\tau_2, t)\psi_{uu}(\tau_1, \tau_2, t)d\tau_2 = \int_{-\infty}^\infty W^*(\tau_2, t)\psi_{us}(\tau_1, \tau_2, t)d\tau_2, \quad \tau_1 \geqslant 0, \quad (60)$$

where $W(\tau, t)$ is the impulsive response of the optimum filter; $W^*(\tau, t)$ is the impulsive response of the filter yielding the desired output; $\psi_{uu}(\tau_1, \tau_2, t)$ is the covariance function of $u(t)$,

$$\psi_{uu}(\tau_1, \tau_2, t) = \langle u_1 u_2 \rangle_{Av}; \quad (61)$$

and $\psi_{us}(\tau_1, \tau_2, t)$ is the covariance function of $u(t)$ and $s(t)$, i.e.,

$$\psi_{us}(\tau_1, \tau_2, t) = \langle u_1 s_2 \rangle_{Av}. \quad (62)$$

[9] The optimization of time-variant linear filters is treated in recent papers by C. L. Dolph and M. A. Woodbury, Trans. Am. Math. Soc. **72**, 519 (1952); R. C. Booton, Proc. Inst. Radio Engrs. **40**, 977 (1952); and R. C. Davis, J. Appl. Phys. **23**, 1047 (1952).

Turning to finite-memory filters, we note that the characteristic function of a filter which is restricted to acting on a finite part of the past of the input, say from $\tau=0$ to $\tau=T$, should vanish outside the interval $0 \leqslant \tau \leqslant T$. As a consequence, the integral equation satisfied by the characteristic function of an optimum finite-memory filter of class $\mathfrak{N}_1$ can be obtained at once from (29) by restricting the range of variation of τ to the interval $(0, T)$. This yields the following integral equation[9]:

$$\int_0^T \int_{-\infty}^\infty K[u_2, \tau_2]p(u_1, u_2; \tau_1 - \tau_2)du_2 d\tau_2 = \int_{-\infty}^\infty \int_{-\infty}^\infty K^*[s_2, \tau_2]p(u_1, s_2; \tau_1 - \tau_2)ds_2 d\tau_2, \quad (63)$$

of which (29) may be regarded as a special case.

Similar modifications in (56) yield integral equations corresponding to (59) and (63) for filters of class $\mathfrak{N}_m$.

[9] An explicit expression for the impulsive response of an optimum finite-memory linear filter has been given by L. A. Zadeh and J. R. Ragazzini, J. Appl. Phys. **20**, 645 (1950).

JOURNAL OF APPLIED PHYSICS VOLUME 24, NUMBER 4 APRIL, 1953

Effects of Very High Pressures on Glass*

P. W. BRIDGMAN, *Harvard University, Cambridge, Massachusetts*

AND

I. ŠIMON, *Arthur D. Little, Inc., Cambridge, Massachusetts*

(Received November 21, 1952)

In contrast to other inorganic solids, glasses may be permanently compacted by application of pressures of the order of 10^4 to 10^5 atmospheres. This effect was studied on two simple oxide glasses (SiO_2 and B_2O_3) and several silicate glasses. The effect of compacting was studied by measuring the densities, dimensions, and x-ray diffraction patterns. A definite threshold pressure is observed in vitreous silica and silicate glasses, under which no effect takes place and above which the collapse takes place readily. Vitreous boric oxide behaves in a different manner, collapsing gradually, starting at the lowest pressures. Vitreous boric oxide exhibits also plastic flow with subsequent strain hardening. X-ray diffraction measurements performed on vitreous silica and boric oxide indicate that the compacting proceeds on the atomic scale, leaving, however, the short-range order of the basic structural units unchanged. The density of compacted glass can be restored to the original value by annealing to sufficiently high temperatures. Activation energy of this process was determined in the case of vitreous boric oxide.

I. INTRODUCTION

IN the region of ordinary pressures (up to approximately 10 000 atmospheres), most glasses behave in a perfectly elastic manner. However, when the pressure is raised close to the 100 000 atmospheres range or beyond, glasses begin to exhibit signs of plastic flow, strain hardening, and permanent changes in their properties. These phenomena have been first observed in connection with an investigation of the effects of pressure up to 100 000 atmospheres on the electrical resistance of elements, alloys, and compounds.[1] Furthermore, since the atomic configuration of a glass is probably not uniquely fixed as is that of a crystal, a permanent change of density in a glass after exposure to hydrostatic pressure is perhaps not unnatural to reflection after the event. However, it has been the almost universal opinion that all substances without mechanical flaws exhibit perfect volume elasticity. It is probable that other types of substances also have a similar imperfect volume elasticity; it has recently been observed in a polymer resin.

The degree of compacting by pressure is continuously variable within certain limits, a fact which opens a possibility of studying any glass of a given composition in different varieties of density, refractivity, etc. It is also conceivable that some properties of compacted glasses, such as mechanical strength, hardness, elastic constants, thermal expansion, etc., are different from those of the original material. These possibilities have not yet been fully explored.

The permanent compacting of glass is temperature dependent to some degree. Although only a few preliminary experiments have been performed so far, it appears that the effect of compacting may become larger when the temperature at which the pressure is applied is increased; this occurs at rather moderate temperatures, far below the softening point of the given glass. Experimentation becomes difficult in this direction because the combination of high pressures and elevated temperatures poses a formidable problem from the point of view of strength of available materials. Therefore, it is not yet known whether the compacting of glass would continue steadily with increasing temperature; it is conceivable that at high temperatures and under high pressure the compacting of the glass phase might change into crystallization with rather sudden segregation of the generally denser, crystalline phase. Otherwise, it is probable that the relative degree of compacting would gradually decrease with increasing temperature because of the opposing effect of thermal agitation.

II. EXPERIMENTAL TECHNIQUES

When working in the range of pressures above 50 000 atmospheres, it becomes difficult to operate under strictly hydrostatic conditions because of the limitations imposed by the mechanical properties of the materials used for the pressure vessels and because of the limited choice of the pressure transmitting media. However, uniaxial pressures of the order of 10^5 atmospheres can be applied on solids in the form of small, thin samples, using the method described in reference 1. A thin disk is compressed between two flat dies made of hard, strong material such as sintered tungsten carbide (Fig. 1). Since the whole load is supported by the specimen itself, the only limitation on the maximum obtainable stress is the compressive strength of the dies. In the arrangement of Fig. 1, the load is confined to a small central area of the die, so that the surrounding material provides support for the most severely stressed region. In addition, the dies are supported on the cylindrical surface by shrunk-on steel rings, providing radial prestressing. In this way, it is possible to attain mean stresses two to three times

* This research was sponsored by the Owens-Illinois Glass Company, Toledo, Ohio. The authors wish to thank the Owens-Illinois Glass Company for permission to publish this paper.

[1] P. W. Bridgman, Proc. Am. Acad. Arts Sci. 81, 165, (1952).

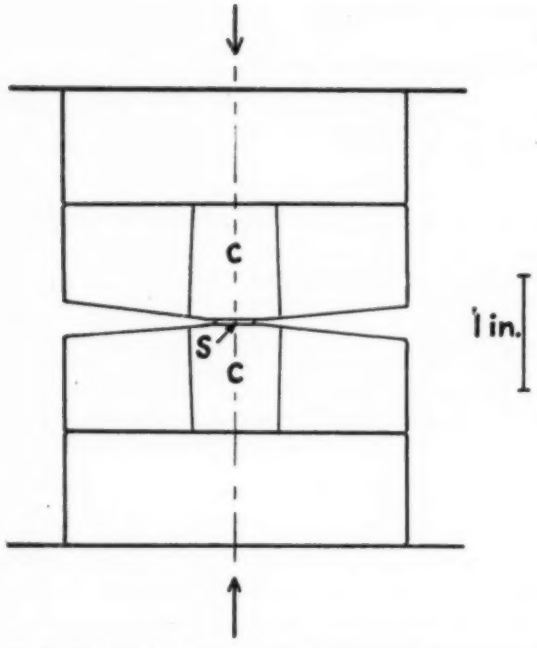


FIG. 1. Schematic diagram of the compression apparatus. S—specimen; C—Carboloy inserts.

higher than the normal crushing strength of sintered tungsten carbide, which is of the order of 60 000 atmospheres. The force is applied through hardened steel end plates by means of a hydraulic press.

The diameter of the flat faces of the dies used in this work was 0.50 inch. The conical angle of faces of the dies outside of the flat area was approximately 170°. The dies were made of Carboloy, Grade 905 (supplied by the Carboloy Company, Detroit, Michigan), ground with a 1 to 50 taper and forced in the hardened steel rings with an excess of 0.010 inch on a 0.780-inch mean diameter. The radial stress on the surface of the Carboloy insert was of the order of 400 000 psi or 27 000 atmospheres. This exceeds the yield strength of the Bethlehem Omega tool steel, from which the ring was made and which was of the order of 280 000 psi, when hardened and tempered to Rockwell-C59. The end plates had the same outside diameter as the rings, and they were made of the same steel. The whole assembly was placed between the platens of a 75-ton hydraulic press.

Although the dies are made of a hard material having very low compressibility (Young's modulus for Carboloy $\approx 90\times 10^6$ psi as compared with $\approx 33\times 10^6$ psi for hardened steel), they deform under the load, and, therefore, the distribution of stress over the area of contact is nonuniform. Moreover, if the diameter of the sample is less than the diameter of the flats on the dies, the compressed specimen is embedded into the originally flat faces of the dies until these eventually cup over and touch each other on the periphery. Before this limit is reached, the mean stress on the sample may be taken as total load divided by the total area of the disk; beyond the limit imposed by the cupping effect, the determination of the actual pressure on the sample becomes almost impossible. Although we have always strived at staying below this limit, it was difficult in the region of extreme pressures to determine when the dies began to come in contact at the periphery. Consequently, the actual mean pressure on the sample might have been less than that which was indicated by the nominal figures. The actual stress in the central area of the sample is also smaller than the nominal value because of its nonuniform distribution. A rigorous calculation of radial distribution of stress over the surface of an elastic disk compressed between two elastic dies demands knowledge of the distribution of strain. This would be difficult to calculate. By an elaborate study of the lenticular shapes of permanently deformed disks of boric oxide glass after compression, there is a possibility that a basis might be obtained for calculation of the distribution of stress over the surface of the disk, but no satisfactory measurements of this kind were made in the present work. One can say, in general, that the normal stress at the center of the disk is less than average and that the radial tangential (shearing) stresses have their largest values near the periphery.

The problem is further complicated by the fact that glass exhibits some degree of plastic flow accompanied by strain hardening, so that variations with time should be also considered. Because of the variability of all these factors, no attempt was made to apply corrections for nonuniform distribution of pressure. Instead, the mean or "nominal" pressures were used throughout this work. The nominal pressure was calculated on the area of the specimen after release of pressure. Some error is introduced here by indefiniteness of the dimensions at the fractured edge. In general, this error probably works in the direction of a diminution of the correct average pressure. The properties of compressed samples were studied mostly on fragments taken from the central area of the disks, where the pressure was undoubtedly lower than the nominal value, in particular at the highest pressures used.

The samples for compression tests were in the form of thin disks, approximately 5 to 8-mm diameter, 0.15 to 0.25-mm thickness, fabricated by slicing, grinding, and polishing from solid stock, and finally cutting out from the slice by means of abrasive-loaded glass tubes. Since each compression to a desired pressure required a new specimen and the number of these was limited by practical considerations, it was not always possible to establish as many points on the measured curves as would be desirable. Most of the silicate glasses and the vitreous boric oxide were prepared in the General Research Laboratories of the Owens-Illinois Glass Company; the compositions indicated are those obtained by analysis of finished melts. Vitreous silica was the commercial product of General Electric Company. This, as well as the samples of Corning

Vycor glass were used in the form of microscope cover slides.

The pressure on the dies was applied by means of a hydraulic press which, with a maximum oil pressure of 1920 psi on 10-inch ram, produced a total thrust of 75 tons = 68 000 kg.

The pressure was applied, as a rule, at a rate of approximately 2 katmos/sec. It was observed that a rapid application of pressure (within a few seconds) produces sometimes appreciably larger effects than with slow application. For instance, on compression of a 90 percent SiO_2–10 percent Na_2O glass to 135 katmos, the relative increase in density was approximately 6 percent on slow compression, while it amounted up to 11 percent on rapid compression. This might have been caused by adiabatic increase of temperature in the sample, which was undesirable; therefore, in all regular runs rapid compression was avoided.

In order to allow for possible flow and associated strain-hardening, the pressure was held constant for one minute after attaining the desired value. This practice was not quite satisfactory in the case of vitreous boric oxide, which can sustain very large plastic deformations. After reaching and holding the top pressure, the load was removed slowly at the same rate as it was applied. It was observed that in the case of "hard," nonplastic glasses, such as vitreous silica, a rapid removal of pressure results in an explosive disintegration of the sample; the "softer," more plastic glasses were more resistant to the sudden release of pressure. The recovery of the samples for examination after compression was greatly facilitated by use of lead rings, approximately 1.5 mm wide, placed around the glass disks. They were cut out of lead foil of approximately the same thickness as the sample and they had a clearance on diameter of approximately 0.5 mm or less. Using this technique, the glass disk could be recovered as a sufficiently compact mass even after its edge became disintegrated.

With the exception of samples compressed to moderate pressures (up to 30 katmos), the glass disks usually came out fractured. Typical kinds of fractures are shown in the photographs in Fig. 2. The hard, nonplastic glasses develop circular, concentric fissures, usually leaving the central part fairly well preserved for subsequent measurements. It is probable that fracturing occurs upon release of pressure when the dies come back elastically to the original shape, which the permanently deformed glass disk cannot follow. In contrast to this, the "soft," plastic glasses, such as vitreous boric oxide, develop radial fissures, sometimes with obvious gaps between the fragments near the periphery. This can be explained by strain-hardening of the material in the peripheral region, where, because of the cupping effect, the material suffers largest reduction in thickness and largest radial deformation. Consequently, the material may become so strong there that it does not sustain enough lateral extension to retain a continuous form. The radially fractured disks were usually dome-shaped after the release of pressure, which indicates the amount of elastically recovered deformation of glass as compared to that of the confining lead ring which is presumably completely plastic.

Although most experiments were performed at ambient temperatures, it was possible to bring the dies to a temperature of up to 200°C by means of electric heaters built into the platens of the press. A few compression tests were performed on vitreous B_2O_3 at 60C° and on vitreous silica at 150°C with no deleterious effects to the dies. The tungsten carbide material itself should be usable to much higher temperatures, although no information is available on its behavior on repeated application of extreme stresses under these conditions. Even at room temperature Carboloy suffers some permanent deformation by repeated application of localized pressures of the order of 100 000 atmos. For this reason, the dies had to be resurfaced occasionally, using diamond lap. We have not observed in these experiments, however, any signs of incipient fractures after a few hundred experimental runs.

III. EXPERIMENTAL RESULTS

Permanent Increase of Density

The densities of the samples before and after compression were obtained by the method of floating the samples in a mixture of two liquids (tetrabromoethane, $d = 2.95$ g/cm³ and carbon tetrachloride, $d = 1.45$ g/cm³) and determining the density of each individual mixture

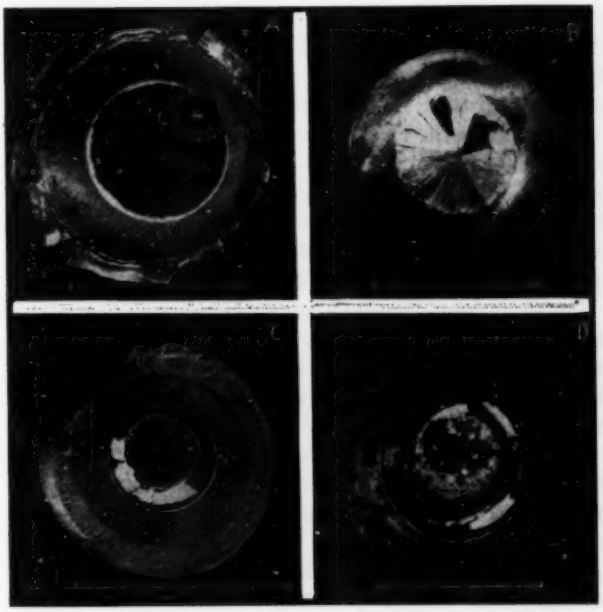


FIG. 2. Fragments of glass disks after compression (some of the disintegrated material was lost on transfer to the photographic camera). A—plate glass, 140 katmos; B—vitreous B_2O_3, 63 katmos at 60°C; C—Vycor glass 100 katmos, D—10 percent Na_2O–90 percent SiO_2 glass, 140 katmos. 1-cm length indicated in each picture.

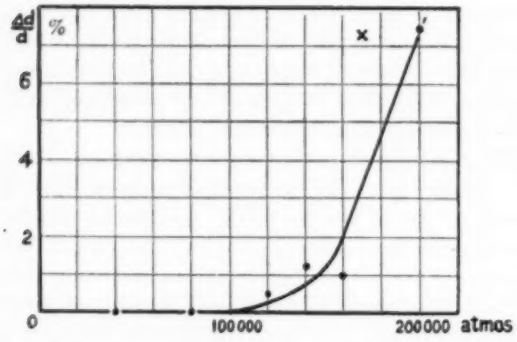


FIG. 3. Relative increase (percent) of density of vitreous silica as a function of applied pressure.

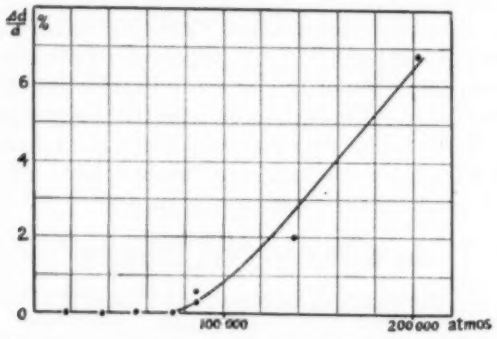


FIG. 4. Relative increase of density of Vycor glass as a function of applied pressure.

by means of a pycnometer. Densities were measured at ambient temperatures (between 20° and 30°C) and corrections were made for the expansion of the liquids and of the pycnometer but not for the samples. Densities could be determined by this technique with an accuracy of approximately ±0.25 percent, which was much better than the accuracy in determining the mean pressure from the applied load and the final dimensions of the sample. Some variability in the results was also caused by the gradual disappearance of the pressure effects caused by "self-annealing" of the compressed samples. This relaxation proceeded rather rapidly in some of the "soft" glasses, in particular vitreous B_2O_3, even at ambient temperatures. The variations in the length of the time interval (10 to 30 minutes) elapsed between the compression and the determination of density might have affected consequently some of the results.

The results of measurements are presented in Figs. 3, 4, 5, and 6 in terms of the relative increase of density $\Delta d/d$ (percent) plotted *versus* the applied nominal pressure P.

Pure vitreous silica (Fig. 3) shows a marked threshold pressure near 100 katmos, below which no permanent effect is observed. After this critical pressure is exceeded, the structure of silica seems to collapse, as indicated by the rapid increase of density to almost 7.5 percent at 200 katmos nominal. The latter figure for the pressure is probably too high in view of the nonuniform distribution of stress on the sample, as discussed in the preceding section; using the actual stress near the center would make the slope of the curve even steeper. One point, obtained on the same material at a temperature of 150°C (marked by a cross in Fig. 3), seems to indicate an appreciable increase of the permanent effect with temperature.

Occasionally, fragments of compacted vitreous silica were obtained with considerably higher densities; one specimen was recovered with density as high as 2.61 g cm^{-3}, as compared with 2.22 g cm^{-3} of the original glass (17.5 percent increase). Although this is quite close to the density of quartz (2.65 g cm^{-3}) and definitely above the densities of cristobalite (2.34 g cm^{-3}) and tridymite (2.33 g cm^{-3}), this specimen remained completely amorphous, as shown by microscopic and x-ray diffraction examination. Unfortunately, corresponding pressures could not be determined, because of the indefinite area of the partially fractured and disintegrated disks. Traces of tungsten carbide abraded from the dies were detected by x-ray analysis on the surface of these highly compacted specimens [Fig. 9(B)]. However, the minute quantities of tungsten carbide actually present cannot account for the observed high densities of the fragments. The samples were still perfectly transparent, and their density could be substantially reduced by annealing in vacuum.

Generally, similar behavior is observed with Vycor glass (Fig. 4), which is a high silica glass manufactured

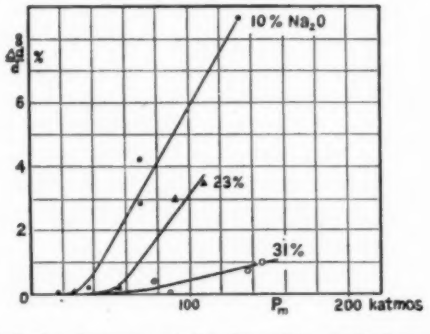


FIG. 5. Relative increase of density with applied pressure for three silicate glasses with 10, 23, and 31 molar percent Na_2O.

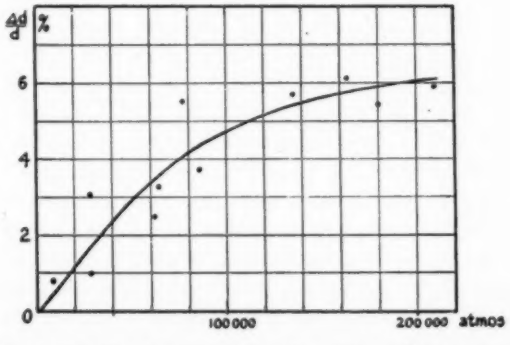


FIG. 6. Relative increase of density with applied pressure for vitreous B_2O_3.

TABLE I.

Molar percentage alkali	Cation radius (A)	P (katmos)	$\Delta d/d$ (percent)
22 Li_2O	0.60	112	6.9
23 Na_2O	0.95	109	3.5
23 K_2O	1.33	128	1.2

by sintering a leached out borosilicate matrix; its composition is given as 96 percent SiO_2, three percent B_2O_3 and one percent other oxides. The threshold pressure is a little lower, near 80 katmos, and the curve has a smaller slope, reaching approximately 6.5 percent permanent increase in density at 200 katmos.

Glasses of the silica-soda family show a progressive degradation of the effects observed with pure silica and 96 percent silica glass (Fig. 5). Only a limited number of samples of the three glasses with molar percentages 10 percent, 23 percent, and 31 percent Na_2O were available, and, therefore, the data are rather incomplete. The threshold pressures are substantially lower, in the vicinity of 40 katmos. At the same time, a marked decrease of the permanent compressibility is observed with increasing content of soda. Obviously, with the increasing number of metallic cations embedded in the interstices of the silica network, a permanent collapse of the structure becomes more difficult.

The kind of cation involved should be also of importance because of the difference in size and in the effective charge. We have compared the permanent increase of density of three silica-alkali glasses having approximately the same molar percentage of alkali oxide at pressures near 100 katmos. The results are shown in Table I. The decrease of the effect with increasing size of the cation is marked.

As a last glass, we have studied vitreous boric oxide (Fig. 6). This glass exhibits much higher permanent change of density in the region of low pressures than any of the other glasses studied before, and there is no observable threshold pressure. This would suggest a difference in the geometrical involvement of the empty spaces responsible for the permanent changes of density. Furthermore, vitreous boric oxide probably has an asymptotic increase of density near 6 percent while there is no evidence in the present pressure range of a corresponding asymptote for the other glasses. This means at least a difference in the scale of magnitude of the forces and may also involve a different mechanism of the process of compacting.

TABLE II.

Molar percentage Na_2O	P (katmos)	$\Delta d/d$ (percent)	$-\Delta h/h$ (percent)	$-\frac{(\Delta h/h)}{(\Delta d/d)}$
10	136	8.7	16	1.8
23	109	3.5	12	3.4
31	138	0.7	25	36

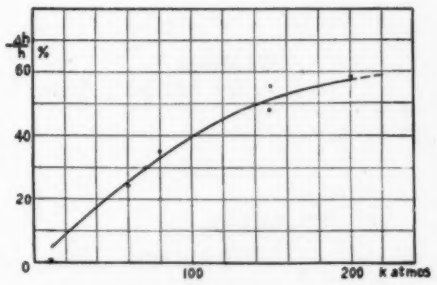


FIG. 7. Relative decrease of thickness of vitreous B_2O_3 disks with applied pressure.

Some additional experiments were performed on compression of vitreous boric oxide under hydrostatic conditions. Two specimens of the same stock were compressed to 40 katmos, using indium as a pressure-transmitting medium. One of the samples, which was kept at a temperature of −79°C during the time between compression and measuring the density, had the density increased by 4.0 percent. The other, kept at ambient temperature for about one hour after compression showed 3.4 percent increase in density. Both figures are larger than the corresponding values for uniaxial compression as shown in Fig. 6, but it is to be recalled that the "nominal" pressures are probably too high.

Permanent Changes of Dimensions

Under uniaxial compression, the thickness of disks made of the "hard" glasses shows a permanent decrease corresponding approximately to the increase of density. For instance, a sample of Vycor glass had its thickness permanently reduced by 7.6 percent after compression to 200 katmos while its density increased by 6.8 percent. The softer glasses always exhibit reductions in thickness larger than the corresponding increase in density. Obviously, lateral expansion is taking place at the same time, making up for excessive decrease of thickness. The very soft glasses (vitreous B_2O_3) extend so much in the lateral direction that the total volume after expansion could be directly obtained by measuring the dimensions. Otherwise, in most cases it was impossible to determine the lateral dimensions with sufficient accuracy because of the disintegration of the edges on compression.

The change in dimensions of a given glass under pressure is significant. It is interesting to observe how the deformation increases in alkali-silicate glasses with increasing percentage of soda (Table II). Most of the total deformation observed in the 31 percent soda glass is of a nonisotropic nature; actually, with its low permanent compressibility and its high flow, this glass behaves more like a plastic than a brittle solid.

The nonisotropic flow is probably most pronounced in vitreous boric oxide. Reductions of thickness observed with vitreous B_2O_3 were as much as 58 percent after compression to 200 katmos (Fig. 7).

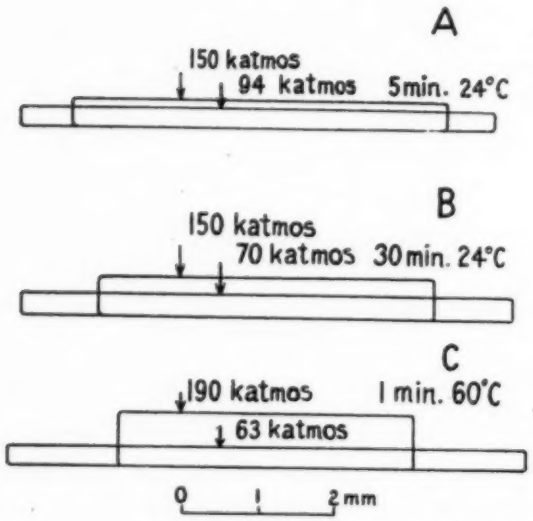


FIG. 8. Cross sections of three disks of vitreous B_2O_3 before and after compression. Initial and final pressures are indicated in the figures. The relative increase of density in sample A was 2.7 percent in B 3.7 percent, and in C 4.7 percent.

The high deformability is here combined with a larger permanent compressibility than with the high soda glasses. If the flow were "ideally plastic," the sample should yield rapidly above a critical pressure until completely extruded out of the dies. A time-dependent asymptotic flow is actually observed. This means that strain-hardening occurs which eventually sets a limit to the lateral extension so that the compressed disk assumes a finite size. This process is rather slow at ambient temperatures and it would probably take hours to complete. A few experiments illustrating this behavior are reproduced in Fig. 8. A B_2O_3 disk, 0.36 mm thick, 4.9 mm in diameter (volume 67.8 mm^3) was compressed with a force of 29 000 kg which was held constant for five minutes. After compression, the thickness was found reduced to 0.25 cm (70 percent of the original thickness) while the diameter increased to 6.2 mm, causing the nominal pressure drop from 150 katmos at the beginning to 94 katmos at the end of the experiment. The volume computed from the dimensions was 68.0 mm^3 after compression; the apparent increase is probably caused by formation of open cracks (see Fig. 2). A thicker disk [Fig. 8(B)] of the same glass (0.55-mm thickness, 4.45-mm diameter) was then compressed with a constant force of 23 000 kg for 30 minutes. The initial nominal pressure was 150 katmos; the final pressure was 70 katmos. Despite larger thickness and lower final pressure, the second sample experienced a larger plastic deformation, its thickness decreasing to 56 percent of the original and the diameter expanding to 6.5 mm, more than the final diameter of the more stressed sample in Fig. 8(A). Finally, in order to observe the temperature dependence of the plastic flow, a third disk of the same B_2O_3 glass was compressed by a constant force of 25 000 kg at 60°C [Fig. 8(C)]. Although the pressure was held constant only for one minute and dropped to 63 katmos, this disk suffered the highest plastic deformation, changing its dimension from 0.70-mm thickness and 3.85-mm diameter to 0.24-mm thickness and 6.8-mm diameter. The accompanying permanent increase of density was 4.7 percent.

X-Ray Analysis of Permanently Compacted Glasses

Structural changes produced in glasses by application of high pressures can be studied by means of the x-ray diffraction method. So far, we have limited ourselves to the two simple oxide glasses, namely, vitreous B_2O_3 and SiO_2; in addition, we have taken data on Vycor glass (96 percent SiO_2), of which we had compressed specimens better suited for x-ray diffraction analysis.

Because of their amorphous structure, glasses produce diffuse patterns when examined by x-ray diffraction. Warren and his co-workers have worked out a method for evaluating these patterns by means of Fourier analysis.[2,3]

This method requires experimental data taken under special precautions, such as strictly monochromatic radiation, removal of air scattering, etc., which would have been too time consuming in the exploratory phase

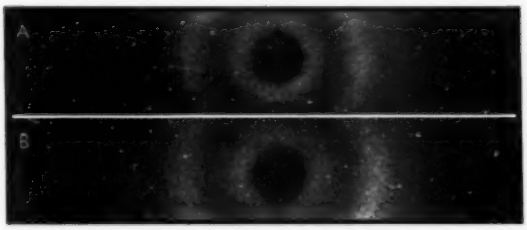


FIG. 9. X-ray diffraction patterns of vitreous silica before compression (A) and after compression to density 17.5 percent higher than original (B). Faint lines visible in (B) are caused by traces of tungsten carbide abraded from the dies.

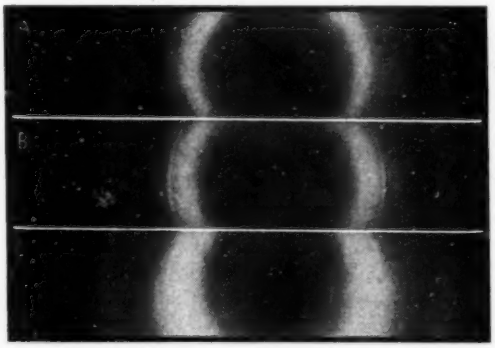


FIG. 10. X-ray diffraction patterns of vitreous boric oxide. A—before compression; B—after compression to 95 katmos; C—after 180 katmos. Sharp lines are caused by a layer of hydrate formed on the surface during exposure.

[2] Warren, Krutter, and Morningstar, J. Am. Ceram. Soc. **19**, 202 (1936).
[3] B. E. Warren, Chem. Revs. **26**, 237 (1940).

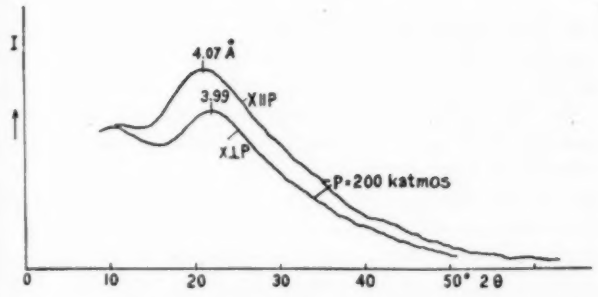


FIG. 11. Anisotropy of Vycor glass compacted by 6.8 percent in density when examined with x-ray beam parallel and perpendicular to the direction of compression.

of this study. Consequently, the data obtained by standard diffraction technique used in this work allow only interpretation of the gross features of the structure. All of the x-ray photographs in this work were taken with filtered Cu K-α radiation, using a 114.7-mm diameter diffraction camera. Only solid, not powdered, specimens were examined and they were not rotated during exposure. Films were evaluated both visually and by means of a recording microphotometer.

Upon x-ray analysis, the specimens of permanently compacted glasses exhibit a distinct uniaxial anisotropy. In general, the diffraction halos of compacted specimens are expanded as compared with the pattern of the original glass. However, this is observed only when the flat specimen is mounted with the direction of compression perpendicular to the incident x-ray beam. This indicates that a decrease in the average interatomic distances takes place predominantly along the direction of applied pressure. Some of the original patterns are reproduced in Figs. 9 and 10. Microphotometer tracings of diffraction halos obtained on compacted Vycor glass are shown in Fig. 11.

In the case of vitreous SiO_2, the strongest halo may be interpreted, at least formally, as caused by diffraction on randomly arranged structural units containing sets of planes with maximum electronic densities separated, on the average, by $d_1=4.11$A. These sets of planes are rudiments of the long-range order prevailing in cristobalite, which is the nearest crystalline substance affined to vitreous silica. In high cristobalite with the edge of the cubic unit cell of 7.12A, the 4.11A spacing is that of the octahedral (111) planes. The observed decrease of d_1 in permanently compacted vitreous silica and Vycor glass does not imply that the Si—O bond lengths themselves are shortened. It is more likely that the separation of the (111) planes is decreased by a change of the bond angles in the interlinked SiO_4 tetrahedra. On this basis, it is conceivable that the existence of a critical pressure (see Figs. 2 and 3) could be explained in a way analogous to Euler's theory of collapse of elastic struts under thrust.

The microphotometer tracings of the x-ray diffraction patterns of vitreous B_2O_3 (Fig. 10) are reproduced in Fig. 12. In addition to the two diffraction halos belonging to vitreous B_2O_3, there are a few sharp lines caused by the crystalline hydrate, which develops in a finely crystalline form on the surface of the sample during exposure in moist air.

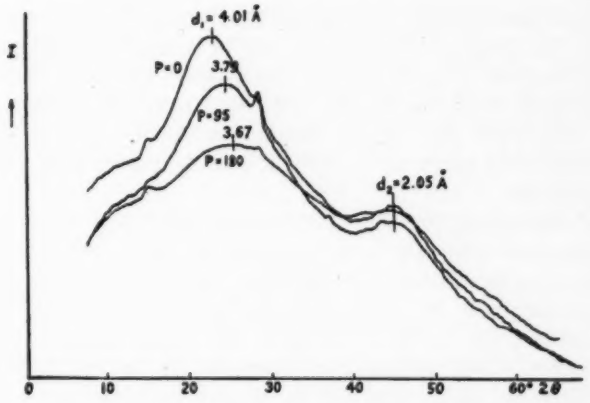


FIG. 12. Microphotometer records of the diffraction patterns of Fig. 10. Note expansion of the d_1 spacing and constancy of the d_2 spacing.

From Warren's analysis,[2] it is known that in vitreous B_2O_3 the average B—O distance is 1.39A and the O—O distance is 2.40A. The structure was described as a random network of interlinked triangular BO_3 units with each oxygen shared by two borons. This allows an interpretation of the observed diffraction halos. The larger halo (d_2) may be related to the size of the oxygen triangle around each boron with side lengths O—O $=2.40$A; the altitude of this triangle is $d_2=2.40\text{A} \times \cos 30° = 2.07$A in agreement with the observed value of 2.05A. The smaller halo (d_1) may be related to the length of units consisting of two BO_3 triangles sharing a common apex; consequently, $d_1=2d_2=4.14$A, in close agreement with the observed figure of 4.16A. It should be pointed out on this occasion that the figures obtained in our work ($d_1=4.01$A and $d_2=2.05$A) differ from the data in reference 2. This discrepancy probably indicates some difference in composition or in the thermal history of the samples. Our samples were used as delivered, without any special heat treatment. The glass has been designated as the low temperature form of vitreous

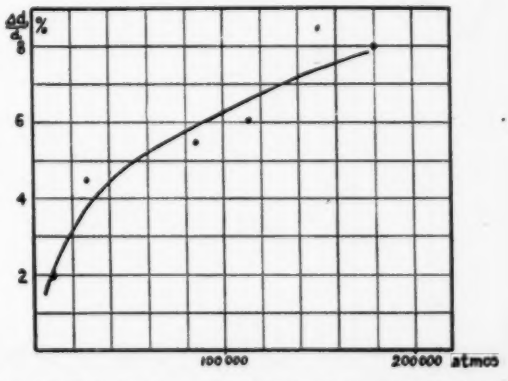


FIG. 13. Relative increase of the lattice spacing d_1 in compacted vitreous B_2O_3 as a function of applied pressure.

B_2O_3 melted at 870°C, cooled slowly to 90°C. However, in the experiments on the effect of temperature on the relaxation of permanently compacted B_2O_3, we have noticed that the density decreased below the original value after prolonged annealing at 300°C in vacuum, indicating incomplete annealing or dehydration of our samples.

Regarding the diffraction patterns in Fig. 12, it is observed that after compression the inner halo expands and becomes broader. The outer halo does not expand, gaining only in width. This indicates that the short-range order of the triangular BO_3 units remains essentially unchanged. The expansion of the inner halo (decrease of d_1 spacing) is caused by some kind of folding up of the structure, probably about the shared oxygens in the neighboring BO_3 groups. The simultaneous broadening of the diffraction halo indicates that this process is not possible without serious distortion of the structure.

Enough data were obtained on vitreous B_2O_3 to plot the dependence of $\Delta d_1/d_1$ on pressure (Fig. 13). It is interesting to observe that the decrease of the spacing d_1 is appreciably larger (by a factor of 1.5) than the corresponding increase in density (Fig. 6). Undoubtedly, there is a compensating effect taking place in a permanent lateral expansion of the structure. Unfortunately, no x-ray data in the perpendicular direction have been taken with this material.

Annealing of Compressed Vitreous Boric Oxide

It can be expected that, in general, all "permanent" effects resulting from compression of glass in the cold state can be removed by heating. The time-temperature relationships governing these changes might provide useful information regarding the nature of the relaxation phenomena. These effects have been studied in the past almost exclusively by means of optical birefringence. With glasses compacted by high pressures, other properties, such as density, atomic interplanar spacings, and dimensional changes can be conveniently studied as functions of time and temperature, in addition to optical birefringence.

We have studied the density-time-temperature relation in compressed vitreous boric oxide, which was well suited for this exploratory study because of the large size of the effect and low annealing temperatures. We have selected for the measurements the specimens of vitreous B_2O_3 which were compressed to 40 katmos under hydrostatic conditions. This was done in order to avoid possible complications by anisotropic effects. Samples were kept and heated in evacuated glass tubes to prevent the formation of hydrate on the surface. Densities were measured by the flotation technique.

The annealing proceeds rather rapidly even at ambient temperatures and, in fact, the density right after removal of the pressure might have been still appreciably higher than what was measured after the necessary handling of the sample. A plot of relative density change as a function of time at two temperatures is presented in Fig. 14. The curves exhibit a sharp drop in the initial period, and then they level off, perhaps tending to an asymptotic limit. The samples were ultimately heated to 310°C for 28 hours in vacuum, resulting in a density of 1.825 g cm^{-3} which was 1.6 percent below the density before compression (1.860 g cm^{-3}), indicating either incomplete annealing or incomplete dehydration of the original glass.

The curves in Fig. 14 do not seem to obey the Adams-Williamson law, nor the exponential law of Maxwell.[4] However, we may make an estimate of the activation energy of the relaxation effect during its initial phase, assuming that it is a rate process with an exponential dependence on temperature. Measuring the rates at two temperatures T_1 and T_2 (K°) and denoting by $r=\Delta d/dt$ the rate of density decrease, we obtain the activation energy as

$$E=k(1/T_1-1/T_2)^{-1}\log(r_2/r_1).$$

Taking for the initial rates the slopes indicated in Fig. 14 by dashed tangents, we obtain $r_2/r_1=1.54$ corresponding to the two temperatures $T_1=298$°K and $T_2=373$°K. From this, we obtain $E=1.3$ kcal/mole for the activation energy of the relaxation of strain in compressed boric oxide. This is a rather crude estimate, being based on a few experimental points taken only at two temperatures. Moreover, the formula is derived under rather oversimplifying assumptions.

A few experiments performed with highly compacted vitreous silica indicated that there does not occur any appreciable self-annealing at room temperature. At elevated temperatures, the density decreases toward the original value. One specimen of vitreous silica, compacted to 17.5 percent density increase, showed a drop in density to 7.9 percent after annealing in vacuum at 430°C for one hour.

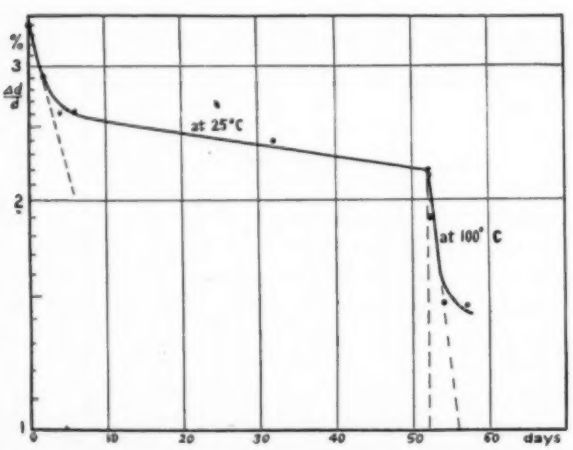


FIG. 14. Annealing curve of compacted vitreous B_2O_3. Relative decrease of density as a function of time at 25°C and at 100°C.

[4] N. W. Taylor, J. Am. Ceram. Soc. **21**, 85 (1938), and also F. W. Cagle and H. Eyring, J. Appl. Phys. **22**, 771 (1951).

IV. CONCLUSIONS

Considering at first vitreous silica and alkali silicate glasses, it appears that, in general, these glasses collapse when the external pressure exceeds some critical value of the order of 50 000 to 100 000 atmos. This effect is most pronounced with pure vitreous silica, and it almost disappears when a large number of sodium ions is incorporated into the silica network.

X-ray examination of compacted vitreous silica and, in particular, of compacted vitreous boric oxide shows that the collapse takes place on the atomic scale leaving, however, the nearest neighbor distance of Si—O or B—O approximately unchanged. Consequently, the compacting of glass may be pictured as some kind of folding up of the vitreous network, presumably involving bending rather than actual shortening of the Si—O or B—O bonds. In case of soda silica glasses the bond bending becomes more difficult when foreign atoms are embedded in the structure.

The folded up structures are mechanically stable at ordinary temperatures, indicating that atoms displaced by bending and otherwise are trapped in the new positions behind fairly strong potential barriers. However, these new positions tend to revert to the original equilibrium state on raising the temperature. In this sense the compacted glasses behave in a way similar to rapidly chilled glasses, which have abnormally low densities. These latter tend to shrink down toward the higher density of a well-annealed glass. These two observations seem to indicate that at any temperature there exists an equilibrium configuration which may be departed from in either direction.

With an increasing amount of added soda to the silica glass and an attendent decrease of permanent compressibility the capacity for permanent deformation with change of shape increases. This can perhaps be explained by an increase in the number of broken Si—O bonds and by degradation of the potential barriers caused by the introduction of the cationic fields into the framework.

In contrast to the silica and silicate glasses, vitreous boric oxide exhibits incomplete volume elasticity even at comparatively low hydrostatic pressures. The compacting seems to reach saturation at pressures of the order of 100 000 atmos.

The Susceptance of a Circular Obstacle to an Incident Dominant Circular-Electric Wave

L. S. SHEINGOLD*
Cruft Laboratory, Harvard University, Cambridge, Massachusetts
(Received September 19, 1952)

The problem of the susceptance of a circular obstacle in a circular wave guide with an incident TE_{01} mode is solved by a variational method. Theoretical expressions are obtained which are in good agreement with experimental results. Curves of normalized susceptance as a function of relative aperture, guide wavelength, and free-space wavelength are included.

I. INTRODUCTION

RECENT interest in the practical transmission of millimeter waves has created a demand for a practical low loss transmission line. The dominant circular-electric mode is ideally suited for this application. During the past few years several researchers have investigated the anomalous attenuation character of this mode and have considered its practical application.[1–6] The problem of exciting a relatively pure TE_{01} mode with sufficiently high available power has been solved satisfactorily by Southworth and his co-workers at the Bell Telephone Laboratories.[7,8] Theoretical and experimental investigations of attenuation in circular wave guide indicate that the TE_{01} mode can propagate with negligible additional attenuation resulting from mode conversion caused by slight random variations in the diameter of standard tubing.[9,10] The theoretical problem of a curved circular wave guide has been considered in great detail.[11,12] The complication arises from the fact that the TE_{01} mode is not the dominant mode in the circular wave guide. In particular the TE_{01} mode has the same phase velocity as the TM_{11} mode, and therefore, any major disturbance such as a curvature of the wave-guide axis will convert the energy of the TE_{01} mode to the TM_{11} mode as well as to other propagating modes. Only uniform circular wave guides are considered here because to date this is the only practical method of transmission of the TE_{01} mode.

This paper is concerned with the determination of the susceptance of a symmetrical obstacle to an incident dominant circular-electric wave. The obstacle shown in Fig. 1 is assumed to possess infinite conductivity and negligible thickness. The obstacle is planar and has circular symmetry. It is assumed that the exciting mode is the dominant circular-electric mode; hence the symmetry of the obstacle limits the higher attenuating modes excited by the obstacle discontinuity to the higher circular-electric modes. The energy stored in the higher modes is entirely magnetic. Accordingly, the circular obstacle can be represented as a shunt inductance on the analogous transmission line.

The general procedure for the determination of the normalized susceptance of a thin obstacle by a variational principle has been described in several references.[13–18] The accuracy of the theoretical results obtained depends on the nature of the assumed trial field. It has been found that the laborious calculations required to evaluate the normalized susceptance usually limit the trial function to a series of two terms.

In this report the problem of the circular obstacle is considered from two aspects, one involving a lower bound variational formulation and the other an upper bound variational formulation.

An expression which is a lower bound to the true value of the obstacle susceptance is expressed in terms of the tangential magnetic field (or current) on the

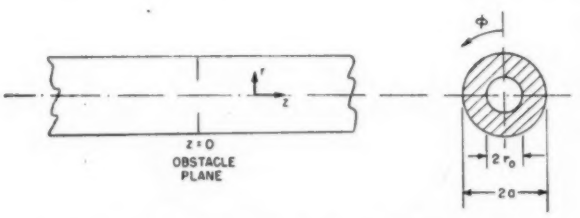


FIG. 1. Obstacle in circular wave guide.

* Now with the Boston Engineering Laboratory of Sylvania Electric Products Inc., Boston, Massachusetts.
[1] G. C. Southworth, Bell System Tech. J. 15, 284 (1936).
[2] L. J. Chu, J. Appl. Phys. 9, 583 (1938).
[3] A. G. Clavier and V. Altovsky, Elec. Commun. 18, 81 (1939).
[4] L. Brillouin, J. Franklin Inst. 229, 709 (1940).
[5] J. Kemp, Wireless Engr. 23, 211 (1946).
[6] J. Ortusi, Ann. Radio Elect. 4, No. 16, 95 (1949).
[7] G. C. Southworth, *Principles and Applications of Waveguide Transmission* (D. Van Nostrand Company, Inc., New York, 1950), pp. 360–366.
[8] M. Aronoff, "Radial Probe Measurements of Mode Conversion in Large, Round Waveguide with TE_{01} Mode Excitation," Paper given at the 1951 Institute for Radio Engineers Convention in New York.
[9] S. P. Morgan, Jr., J. Appl. Phys. 21, 329 (1950).
[10] S. E. Miller and A. C. Beck, "Low-Loss Waveguide Transmission," Paper given at the 1951 Institute for Radio Engineers Convention in New York.
[11] M. Jouget, Cables and Transmission, Paris 2, 257 (1948).
[12] W. J. Albersheim, Bell System Tech. J. 28, 1 (1949).
[13] D. S. Saxon, "Notes on Lectures on Discontinuities in Waveguides," Julian Schwinger, Informal Report of the M. I. T. Radiation Laboratory.
[14] J. W. Miles, Proc. Inst. Radio Engrs. 34, 728 (1946).
[15] N. Marcuvitz, Proc. Inst. Radio Engrs. 36, 728 (1948).
[16] K. L. Dunning and R. G. Fellers, J. Appl. Phys. 22, 1316 (1951).
[17] L. Lewin, *Advanced Theory of Waveguides* (Iliffe and Sons, Ltd., London, England, 1951).
[18] N. Marcuvitz, *Waveguide Handbook* (McGraw-Hill Book Company, Inc., New York, 1951), Vol. 10, M. I. T. Radiation Laboratory Series, pp. 138–167.

obstacle. The chosen trial function consists of two terms: the first is the unperturbed current distribution on an obstacle with no aperture; the second is a correction term. This lower bound expression yields susceptances which are accurate for extremely small apertures.

The normalized susceptance of the circular obstacle which serves as an upper bound to the true value of the susceptance is a function of the tangential electric field in the obstacle aperture. Two different trial functions are used in the upper bound formulation. The first assumed trial function is a series of two terms consisting of the electric field in the aperture, as the radius of the aperture approaches that of the waveguide, and a correction term. Theoretical values for the susceptance obtained by this approach yield accurate results for obstacles with large apertures. The second trial function consists of a two-term trial function for the aperture electric field and is chosen to satisfy the boundary conditions at the obstacle discontinuity in the proper fashion.

Theoretical results obtained from the lower and upper bound expressions are in excellent agreement with the experimental results.

II. VARIATIONAL FORMULATION

The electromagnetic nature of the obstacle problem is simplified by representing the discontinuity as a lumped constant equivalent parameter in a TE_{01} transmission line. This is accomplished by expressing the field components in terms of mode voltages and mode currents. Owing to the symmetrical excitation and the symmetry of the obstacle, the modes of interest are the circular-electric modes. For harmonic time dependence ($\exp(j\omega t)$), the field components of the circular-electric modes expressed as a function of mode voltages and currents are

$$E_r=0,$$

$$E_\phi=\frac{J_1(\rho_n r/a)}{(\pi)^{\frac{1}{2}}aJ_0(\rho_n)}V_n(z)=e_n(r)V_n(z),$$

$$E_z=0,$$

$$H_r=\frac{-J_1(\rho_n r/a)}{(\pi)^{\frac{1}{2}}aJ_0(\rho_n)}I_n(z)=h_n(r)I_n(z), \qquad (1)$$

$$H_\phi=0,$$

$$H_z=\frac{-j\eta\rho_n J_0(\rho_n r/a)}{ka^2(\pi)^{\frac{1}{2}}J_0(\rho_n)}V_n(z),$$

$$e_n(r)=-h_n(r).$$

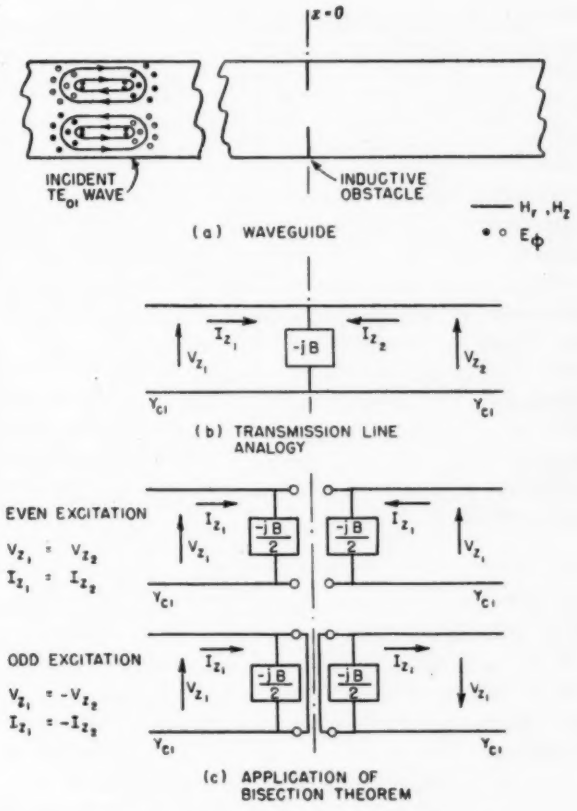


FIG. 2. Simplification of wave-guide circuit.

A tabulation of the necessary quantities is given below.

r_0 = obstacle aperture radius,
a = radius of the circular wave guide,
$\delta=r_0/a$ = relative aperture or aperture radius to guide radius ratio,
$k=\omega(\mu\epsilon)^{\frac{1}{2}}$ = free-space wave number,
ρ_n = nth root of $J_1(x)=0$,
$\beta_n=[k^2-(\rho_n/a)^2]^{\frac{1}{2}}$ = wave number of guide,
$\eta=(\epsilon/\mu)^{\frac{1}{2}}$ = intrinsic admittance of free space,
$\zeta=1/\eta=(\mu/\epsilon)^{\frac{1}{2}}$ = intrinsic impedance of free space.

Substitution of the field components of Eq. (1) into Maxwell's field equations results in two simultaneous equations which are identical in form to the conventional transmission-line equations.

$$dV_n(z)/dz=-jk\zeta I_n(z)=-j\beta_n Z_{cn}I_n(z), \qquad (2a)$$

$$dI_n(z)/dz=(-j\beta_n{}^2/k)\eta V_n(z)=-j\beta_n Y_{cn}V_n(z). \qquad (2b)$$

Using the transmission-line analog, it is possible to define a characteristic admittance and a characteristic impedance for each circular-electric transmission line. For propagating modes the characteristic admittance Y_{cn} and the characteristic impedance Z_{cn} are real quantities, and related by

$$Y_{cn}=\frac{1}{Z_{cn}}=\frac{[k^2-(\rho_n)^2/a]^{\frac{1}{2}}}{\omega\mu}, \qquad (3)$$

whereas for nonpropagating modes Y_{cn} and Z_{cn} are imaginary. Therefore,

$$Y_{cn}=\frac{1}{Z_{cn}}=\frac{-j[(\rho_n/a)^2-k^2]^{\frac{1}{2}}}{\omega\mu}. \tag{4}$$

The thin circular obstacle can be represented as an equivalent shunt susceptance B on the TE_{01} transmission line. The TE_{01} field exciting the obstacle is arbitrary, making it possible to separate the exciting field into even and odd components. The even case requires the tangential magnetic field to vanish in the aperture while the tangential electric field is at its maximum value. Odd excitation requires the tangential electric field to vanish in the aperture and the tangential magnetic field to be maximum. Considering only even excitation and applying[19] the bisection theorem immediately leads to

$$Y_{in}=-j(B/2), \tag{5}$$

where Y_{in} is the admittance looking into the bisected structure (see Fig. 2).

The obstacle susceptance can now be determined by solving Maxwell's field equations in the region to the left of the obstacle with the following boundary conditions:

$$\begin{aligned} &E_\phi=0 \quad r=a \quad -\infty<z<0,\\ &E_\phi=0 \quad r_0<r<a \quad z=0,\\ &H_r=0 \quad 0<r<r_0 \quad z=0. \end{aligned} \tag{6}$$

The problem can be formulated in terms of the tangential electric field in the aperture or the tangential magnetic field on the surface of the obstacle.

The uniform field structure in the ϕ-direction permits the transverse electric field to be expressed as a sum of mode voltages of the circular-electric type. A similar representation in terms of the mode currents can be obtained for the transverse magnetic field. From Eq. (1) these quantities are given by

$$E_t=E_\phi(r,z)=\sum_{n=1}^{\infty}e_n(r)V_n(z), \tag{7a}$$

$$H_t=H_r(r,z)=\sum_{n=1}^{\infty}h_n(r)I_n(z). \tag{7b}$$

The orthogonality conditions permit the mode voltage amplitudes and mode current amplitudes to be expressed in the form of integrals over the cross section of the circular guide. The results are

$$V_n(z)=\int_S E_\phi(r,z)e_n(r)d\sigma, \tag{8a}$$

$$I_n(z)=\int_S H_r(r,z)h_n(r)d\sigma. \tag{8b}$$

[19] A. C. Bartlett, *The Theory of Electrical Artificial Lines and Filters* (John Wiley and Sons Inc., New York, 1930), p. 28.

The application of the boundary condition that the transverse magnetic field must vanish on the aperture of the obstacle yields the relationship

$$I_1h_1(r)=-\sum_{n=2}^{\infty}I_nh_n(r) \qquad \begin{matrix} z=0.\\ 0<r<r_0 \end{matrix} \tag{9}$$

The left side of this equation represents the propagating mode (dominant circular-electric mode) while the summation consists of higher mode terms excited by the discontinuity of the obstacle. These higher nonpropagating modes are assumed to be attenuated completely before reaching the nearest discontinuity. This is equivalent to terminating each nonpropagating mode in its reactive characteristic admittance. With the transmission-line equation relating V_n and I_n, Eq. (9) can be written as

$$\begin{aligned} I_1h_1(r)&=\sum_{n=2}^{\infty}Y_{cn}V_nh_n(r)\\ &=-\sum_{n=2}^{\infty}Y_{cn}h_n(r)\int_{\text{aperture}}E_\phi(r')h_n(r')d\sigma'. \end{aligned} \tag{10}$$

Interchanging the order of summation and integration yields

$$I_1h_1(r)=\int_{\text{aperture}}G(r,r')E_\phi(r')d\sigma', \tag{11}$$

where the function $G(r, r')$ is defined as

$$G(r,r')=-\sum_{n=2}^{\infty}Y_{cn}h_n(r)h_n(r'). \tag{12}$$

Evidently (11) is an integral equation in which the aperture electric field is the unknown quantity. If it is possible to determine $E_\phi(r')$ from the solution of the integral equation, the desired susceptance can be obtained from Eq. (5) and Eq. (8a). It then follows that the susceptance of the obstacle is

$$\frac{jB}{2}=\frac{I_1}{\int_{\text{aperture}}E_\phi(r)h_1(r)d\sigma}. \tag{13}$$

Because of the inherent difficulty involved in obtaining a solution for the integral equation, an approximate expression for the obstacle susceptance will be obtained by the application of a variational principle.

Multiplying Eq. (10) by the unknown aperture field $E_\phi(r)$, integrating over the aperture, and dividing by V_1^2 yields the variational expression for the obstacle

susceptance. It is

$$\frac{-jB}{2}=\frac{\sum_{2}^{\infty} Y_{en}\left[\int_{\text{aperture}} E_\phi(r)h_n(r)d\sigma\right]^2}{\left[\int_{\text{aperture}} E_\phi(r)h_1(r)d\sigma\right]^2}. \quad (14)$$

It can be shown readily that this expression for the obstacle susceptance is stationary with respect to small variations of $E_\phi(r)$ about the correct electric field in the aperture. If the chosen trial function is correct to the first order, the resultant susceptance will be correct to the second order. If the chosen trial field is the *actual* aperture electric field, then the calculated susceptance is a minimum. This expression is in a particularly convenient form because the results are independent of the amplitude of the trial function.

The susceptance determined by the above method is an upper bound to the correct susceptance. A lower bound can be derived from a similar formulation involving an assumed current distribution on the obstacle. The lower bound susceptance obtained in this manner is

$$\frac{-jB}{2}=\frac{\left[\int_{\text{obstacle}} H_r(r)e_1(r)d\sigma\right]^2}{\sum_{2}^{\infty} Z_{en}\left[\int_{\text{obstacle}} H_r(r)e_n(r)d\sigma\right]^2}. \quad (15)$$

III. LOWER BOUND VALUES OF OBSTACLE SUSCEPTANCE

The assumed obstacle current (the tangential component of the magnetic field) consists of two terms; the first term is the obstacle current as the aperture tends to zero and the second term is a correction term. The trial field chosen, consisting of the first two terms of a set of Bessel functions, is

$$H_r(r)=J_1[\rho_1(r/a)]+AJ_1[\rho_2(r/a)], \quad r_0\leq r\leq a, \quad (16)$$

where A is a constant to be evaluated later.

It would be possible to obtain a very accurate solution by assuming as a trial function the sum of a large number of such functions with undetermined coefficients. The calculated results obtained by determining the unknown coefficients so as to make B stationary would yield an accurate value for the susceptance of the obstacle. Practically, however, the calculations become extremely tedious for trial functions consisting of more than two terms. For the practical evaluation of the obstacle susceptance, the analysis of the lower and upper bound expressions is specialized to trial functions containing only two terms.

The assumed field $H_r(r)$ is inserted in Eq. (15), and the variational expression for the obstacle susceptance is written in the following form:

$$\frac{Ba}{Y_{c1}\lambda_g}=\frac{\left[\int_{r_0}^{a} J_1\left(\rho_1\frac{r}{a}\right)\left\{J_1\left(\rho_1\frac{r}{a}\right)+AJ_1\left(\rho_2\frac{r}{a}\right)\right\}rdr\right]^2}{\pi J_0^2(\rho_1)\sum_{n=2}^{\infty}\frac{\left[\int_{r_0}^{a} J_1\left(\rho_n\frac{r}{a}\right)\left\{J_1\left(\rho_1\frac{r}{a}\right)+AJ_1\left(\rho_2\frac{r}{a}\right)\right\}rdr\right]^2}{[\rho_n^2-(ka)^2]^{\frac{1}{2}}J_0^2(\rho_n)}}. \quad (17)$$

The integrals in the previous equation can be evaluated by using these standard forms for the Bessel functions,[20]

$$\int_{r_0}^{a} J_1\left(\rho_m\frac{r}{a}\right)J_1\left(\rho_n\frac{r}{a}\right)rdr=\frac{a^2\delta}{\rho_m^2-\rho_n^2} \qquad m\neq n$$
$$\times[\rho_m J_0(\rho_m\delta)J_1(\rho_n\delta)-\rho_n J_1(\rho_m\delta)J_0(\rho_n\delta)], \quad (18)$$

and

$$\int_{r_0}^{a} J_1\left(\rho_m\frac{r}{a}\right)^2 rdr=\frac{a^2\delta}{2}\left[\delta\{J_0(\rho_m\delta)J_2(\rho_m\delta)-J_1^2(\rho_m\delta)\}-\frac{J_0(\rho_m)J_2(\rho_m)}{\delta}\right]. \quad (19)$$

[20] N. W. McLachlan, *Bessel Functions for Engineers* (Oxford University Press, London, 1934), p. 160.

Equation (17) can be simplified by letting

$$a^2\delta\{G_1+AG_2\}=\int_{r_0}^{a} J_1\left(\rho_n\frac{r}{a}\right)J_1\left(\rho_1\frac{r}{a}\right)rdr$$
$$+A\int_{r_0}^{a} J_1\left(\rho_n\frac{r}{a}\right)J_1\left(\rho_2\frac{r}{a}\right)rdr \quad (20)$$

and

$$a^2\delta\{F_1+AF_2\}=\int_{r_0}^{a}\left[J_1\left(\rho_1\frac{r}{a}\right)\right]^2 rdr$$
$$+A\int_{r_0}^{a} J_1\left(\rho_1\frac{r}{a}\right)J_1\left(\rho_2\frac{r}{a}\right)rdr. \quad (21)$$

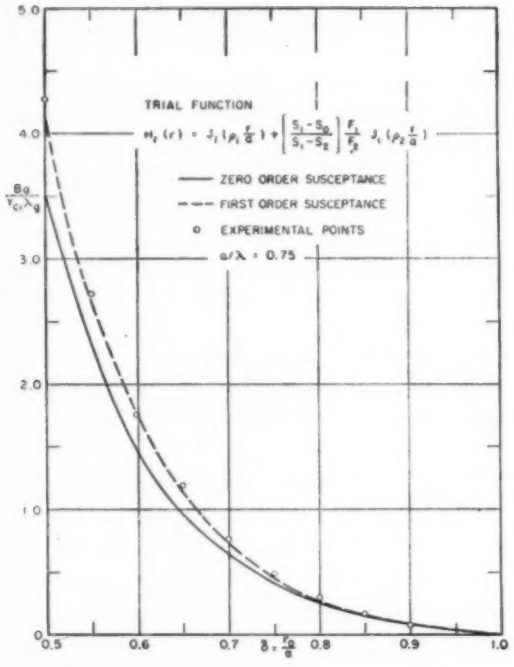


FIG. 3. Lower bound values of susceptance.

Performing the required integrations results in

$$G_1=\frac{[\rho_n J_0(\rho_n\delta)J_1(\rho_1\delta)-\rho_1 J_1(\rho_n\delta)J_0(\rho_1\delta)]}{(\rho_n^2-\rho_1^2)} \quad n\geqq 2, \quad (22)$$

$$\begin{cases} G_2=\tfrac{1}{2}[\delta\{J_0(\rho_2\delta)J_2(\rho_2\delta)-J_1^2(\rho_2\delta)\} \\ \qquad -\{J_0(\rho_2)J_2(\rho_2)/\delta\}] \quad n=2 \\ G_2=\dfrac{[\rho_n J_0(\rho_n\xi)J_1(\rho_2\delta)-\rho_2 J_1(\rho_n\delta)J_0(\rho_2\delta)]}{(\rho_n^2-\rho_2^2)} \quad n>2, \end{cases} \quad (23)$$

$$F_1=\tfrac{1}{2}[\delta\{J_0(\rho_1\delta)J_2(\rho_1\delta)-J_1^2(\rho_1\delta)\} -\{J_0(\rho_1)J_2(\rho_1)/\delta\}], \quad (24)$$

$$F_2=\frac{[\rho_2 J_0(\rho_2\delta)J_1(\rho_1\delta)-\rho_1 J_1(\rho_2\delta)J_0(\rho_1\delta)].}{(\rho_2^2-\rho_1^2)}. \quad (25)$$

Substitution of the above expressions into the equation for the obstacle susceptance yields

$$\frac{Ba}{Y_{c1}\lambda_g}=\frac{[1+A']^2}{\pi J_0^2(\rho_1)\sum_2^\infty \dfrac{\left[\dfrac{G_1}{F_1}+A'\dfrac{G_2}{F_2}\right]^2}{[\rho_n^2-(ka)^2]^{\frac{1}{2}}J_0^2(\rho_n)}}, \quad (26)$$

where

$$A'=A(F_2/F_1) \quad (27)$$

is a constant to be evaluated.

A further simplification is achieved by defining

$$S_0=\pi J_0^2(\rho_1)\sum_2^\infty \frac{[G_1/F_1]^2}{[\rho_n^2-(ka)^2]^{\frac{1}{2}}J_0^2(\rho_n)}, \quad (28)$$

$$S_1=\pi J_0^2(\rho_1)\sum_2^\infty \frac{[G_1G_2/F_1F_2]}{[\rho_n^2-(ka)^2]^{\frac{1}{2}}J_0^2(\rho_n)}, \quad (29)$$

$$S_2=\pi J_0^2(\rho_1)\sum_2^\infty \frac{[G_2/F_2]^2}{[\rho_n^2-(ka)^2]^{\frac{1}{2}}J_0^2(\rho_n)}. \quad (30)$$

The normalized susceptance for the circular obstacle can be expressed simply as

$$\frac{Ba}{Y_{c1}\lambda_g}=\frac{[1+A']^2}{S_0+2A'S_1+A'^2S_2}. \quad (31)$$

The condition for the susceptance to be stationary with respect to the arbitrary coefficient A' is that the derivative of B, with respect to A', must be zero. Performing the differentiation and setting the result equal to zero, one obtains

$$A'=(S_1-S_0)/(S_1-S_2). \quad (32)$$

Inserting the value for A' into Eq. (31) gives

$$\frac{Ba}{Y_{c1}\lambda_g}=1\Big/S_0-\frac{(S_1-S_0)^2}{(S_0+S_2-2S_1)} \quad \text{first order} \quad (33)$$

for the two-term trial function. The susceptance using

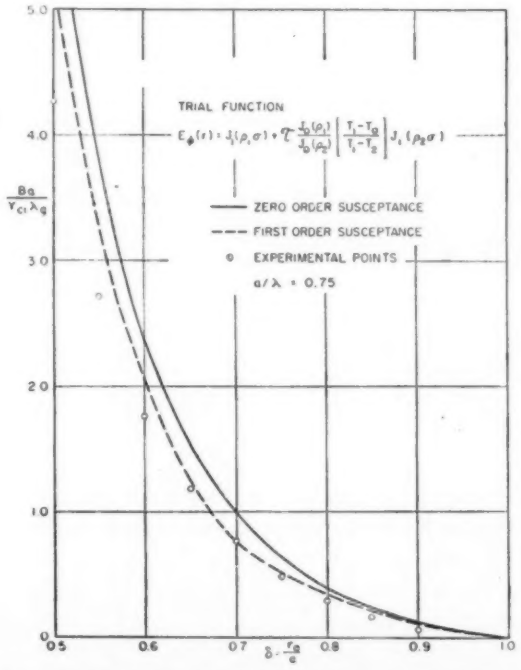


FIG. 4. Upper bound values of susceptance, first trial function.

a one-term trial function ($A'=0$) is given by

$$Ba/Y_{c1}\lambda_g = 1/S_0. \quad \text{zero order.} \tag{34}$$

It can be shown readily that

$$(S_1-S_0)^2/(S_0+S_2-2S_1)$$

in the denominator of Eq. (33) is a positive quantity; therefore the susceptance determined by insertion of the two-term trial function is larger than the susceptance determined from the first term.

Referring to the trial function of Eq. (16), the coefficient A is evaluated to give the maximum susceptance for the chosen trial function, yielding

$$H_r(r) = J_1\left(\rho_1 \frac{r}{a}\right) + \left[\frac{S_1-S_0}{S_1-S_2}\right]\frac{F_1}{F_2} J_1\left(\rho_2 \frac{r}{a}\right). \tag{35}$$

IV. UPPER BOUND EXPRESSION FOR THE OBSTACLE SUSCEPTANCE

Two different trial functions are used in the upper bound variational formulation for the normalized susceptance of the circular obstacle.

The first trial function consists of the unperturbed aperture electric field and a correction term. This assumed field which will give accurate values for the susceptance of obstacles with large apertures (i.e., small perturbation of the incident field) is

$$E_\phi(r) = J_1[\rho_1(r/r_0)] + CJ_1[\rho_2(r/r_0)], \quad 0 \le r \le r_0. \tag{36}$$

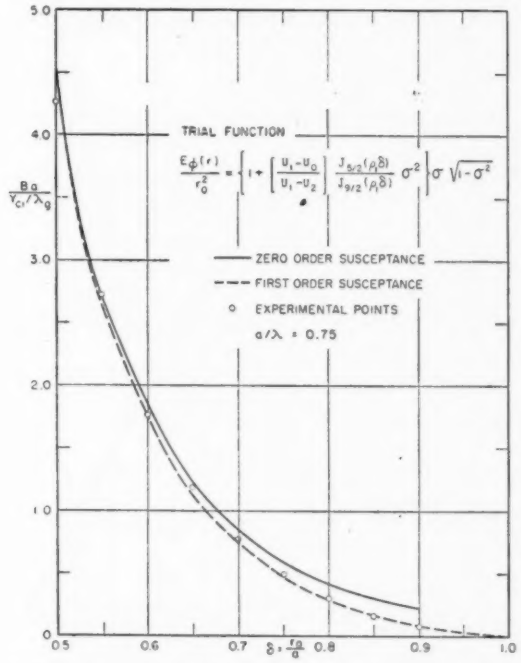


FIG. 5. Upper bound values of susceptance, second trial function.

The above function satisfies the boundary condition that the electric field must vanish at the axis and at $r=r_0$.

Insertion of the trial function into the upper bound expression for the obstacle susceptance allows the variational form for the susceptance to be expressed as

$$\frac{Ba}{Y_{c1}\lambda_g} = \frac{J_0^2(\rho_1)\sum_{n=2}^{\infty}[\rho_n^2-(ka)^2]^{\frac{1}{2}}\left[\int_0^{r_0} J_1\left(\rho_n \frac{r}{a}\right)\left\{J_1\left(\rho_1 \frac{r}{r_0}\right)+CJ_1\left(\rho_2 \frac{r}{r_0}\right)\right\} r\,dr\right]^2}{\pi\left[\int_0^{r_0} J_1\left(\rho_1 \frac{r}{a}\right)\left\{J_1\left(\rho_1 \frac{r}{r_0}\right)+CJ_1\left(\rho_2 \frac{r}{r_0}\right)\right\} r\,dr\right]^2}. \tag{37}$$

The integrations are carried out by Lommel's integral formulas [as in Eqs. (18) and (19)]. The relative susceptance can be simply expressed as

$$Ba/Y_{c1}\lambda_g = (T_0+2C'T_1+C'^2T_2)/(1+C')^2, \tag{38}$$

where

$$T_K = \frac{(\delta^2-1)^{(2-K)}(\delta^2-\tau^2)^K J_0^2(\rho_1)}{\pi J_1^2(\rho_1\delta)} \times \sum_{n=2}^{\infty} \frac{[\rho_n^2-(ka)^2]^{\frac{1}{2}} J_1^2(\rho_n\delta)}{J_0^2(\rho_n)(\alpha_n^2-1)^{2-K}(\alpha_n^2-\tau^2)^K}$$

$$K=0, 1, 2 \tag{39}$$

TABLE I. Lower bound and upper bound values of obstacle susceptance.

	Lower bound		$a/\lambda=0.75$ Upper bound				
δ	$1/S_0$	$S_0-\frac{(S_1-S_0)^2}{S_0+S_2-2S_1}$	T_0	$T_0-\frac{(T_1-T_2)^2}{T_0+T_2-2T_1}$	U_0	$U_0-\frac{(U_1-U_0)^2}{U_0+U_2-2U_1}$	Experimental values
0.50	3.49	4.13	6.21	5.40	4.52	4.48	4.27
0.55	2.33	2.66	3.60	3.10	2.72	2.66	2.71
0.60	1.47	1.75	2.41	2.11	1.86	1.77	1.76
0.65	0.958	1.12	1.51	1.33	1.23	1.12	1.19
0.70	0.649	0.739	1.01	0.890	0.867	0.754	0.767
0.75	0.415	0.450	0.632	0.561	0.592	0.472	0.490
0.80	0.259	0.263	0.398	0.353	0.417	0.290	0.298
0.85	0.142	0.142	0.222	0.201	0.306	0.161	0.163
0.90	0.068	0.071	0.109	0.103	0.223	0.081	0.070

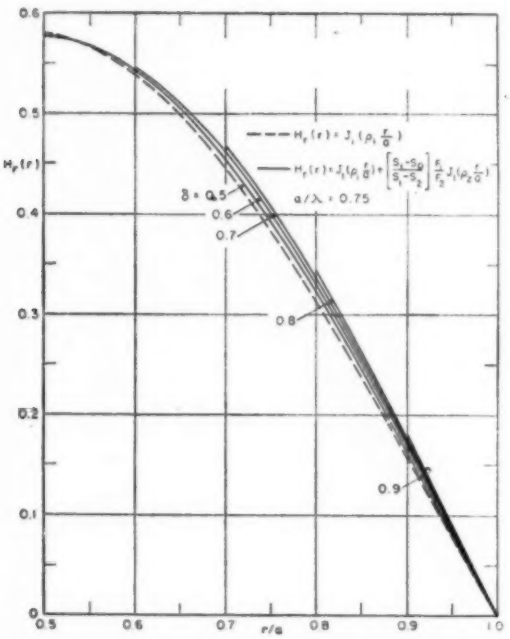


FIG. 6. Assumed current distribution on obstacle.

and the following quantities are defined:

$$C' = CJ_0(\rho_2)/\tau J_0(\rho_1), \tag{40}$$

$$\tau = \rho_2/\rho_1, \tag{41}$$

$$\alpha_n = (\rho_n/\rho_1)\delta. \tag{42}$$

The coefficient C' which yields the minimum susceptance for the chosen trial function is given by

$$C' = (T_1 - T_0)/(T_1 - T_2). \tag{43}$$

Substituting the above quantity into Eq. (38) gives

$$Ba/Y_{c1}\lambda_g = T_0 - (T_1 - T_0)^2/(T_0 + T_2 - 2T_1) \quad \text{first order} \tag{44}$$

for the assumed two-term trial function.

The inductive susceptance determined by using the unperturbed field $E_\phi(r) = J_1[\rho_1(r/r_0)]$ as a trial function is given by

$$Ba/Y_{c1}\lambda_g = T_0 \quad \text{zero order.} \tag{45}$$

It can furthermore be shown that $(T_1 - T_0)^2/(T_0 + T_2 - 2T_1)$ is always a positive quantity indicating that the first-order susceptance of Eq. (44) is less than the zero-order susceptance of Eq. (45).

The trial function of Eq. (36) with C evaluated to give the minimum susceptance for the choice of trial function then becomes

$$E_\phi(r) = J_1(\rho_1\sigma) + \left[\frac{T_1 - T_0}{T_1 - T_2}\right]\tau\frac{J_0(\rho_1)}{J_0(\rho_2)}J_1(\rho_2\sigma), \tag{46}$$

where $\sigma \equiv r/r_0$.

Up to this point trial functions have been considered which are expected to give accurate values for the susceptance of obstacles with very small or very large apertures. The final trial function to be inserted in the upper bound expression of Eq. (14) is an approximation to the aperture electric field determined from the quasi-static solution. The particular function selected is chosen so that the integrals in the variational expression can be evaluated readily.

The trial function which meets these requirements and satisfies the boundary conditions is

$$E_\phi(r) = r(r_0^2 - r^2)^{\frac{1}{2}} + Dr^3(r_0^2 - r^2)^{\frac{1}{2}}. \tag{47}$$

Inserting this function into the variational form for the relative susceptance results in

$$\frac{Ba}{Y_{c1}\lambda_g} = \frac{J_0^2(\rho_1)\sum_{n=2}^{\infty}[\rho_n^2 - (ka)^2]^{\frac{1}{2}}\left[\int_0^{r_0} J_1\left(\rho_n\frac{r}{a}\right)\{r^2(r_0^2 - r^2)^{\frac{1}{2}}[1 + Dr^2]\}dr\right]^2}{\pi\left[\int_0^{r_0} J_1\left(\rho_1\frac{r}{a}\right)\{r^2(r_0^2 - r^2)^{\frac{1}{2}}[1 + Dr^2]\}dr\right]^2}. \tag{48}$$

The integrals to be evaluated are of the form

$$I = \int_0^{r_0} J_1\left(\rho_n\frac{r}{a}\right) r^m(r_0^2 - r^2)^{\frac{1}{2}}dr. \tag{49}$$

These integrals can be evaluated by making the substitution $r = r_0\sin\theta$; it then follows that

$$I = r_0^{(m+2)}\int_0^{\pi/2} J_1(\rho_n\delta\sin\theta)\sin^m\theta\cos^2\theta d\theta. \tag{50}$$

This integral is evaluated from Sonine's first integral formula,[21] yielding

$$I = r_0^{(m+2)}\int_0^{\pi/2} J_1(\rho_n\delta\sin\theta)\sin^m\theta\cos^2\theta d\theta = \frac{\sqrt{2}r_0^{(m+2)}\Gamma(\frac{3}{2})}{(\rho_n\delta)^{\frac{1}{2}}}J_{(m+\frac{1}{2})}(\rho_n\delta), \tag{51}$$

where $\Gamma(x)$ is the gamma-function of x.

Performing the integrations in Eq. (48) the expression for the normalized susceptance becomes

$$Ba/Y_{c1}\lambda_g = (U_0 + 2D'U_1 + D'^2U_2)/(1 + D')^2, \tag{52}$$

where D' is an arbitrary constant defined by

$$D' = Dr_0^2\left[\frac{J_{9/2}(\rho_1\delta)}{J_{5/2}(\rho_1\delta)}\right] \tag{53}$$

[21] N. W. McLachlan, *Bessel Functions for Engineers* (Oxford University Press, London, 1939), p. 159.

and

$$U_K=\frac{J_0^2(\rho_1)(\rho_1)^3[J_{5/2}(\rho_1\delta)]^K}{\pi[J_{9/2}(\rho_1\delta)]^K}$$

$$\times\sum_2^\infty \frac{[\rho_n^2-(ka)^2]^3[J_{5/2}(\rho_n\delta)]^{2-K}}{J_0^2(\rho_n)\rho_n^3[J_{5/2}(\rho_1\delta)]^{2-K}} \quad K=0, 1, 2. \quad (54)$$

Proceeding as in the previous two cases, the arbitrary coefficient is determined by differentiating Eq. (52) with respect to D' and by setting the resultant expression equal to zero. The constant is therefore evaluated to be

$$D'=(U_1-U_0)/(U_1-U_2). \quad (55)$$

It is now possible to express the first-order susceptance as

$$\frac{Ba}{Y_{c1}\lambda_g}=U_0-\frac{(U_1-U_0)^2}{U_0+U_2-2U_1} \quad (56)$$

and the zero-order susceptance as

$$Ba/Y_{c1}\lambda_g=U_0. \quad (57)$$

The trial function of Eq. (47) with the coefficient D adjusted to yield a minimum susceptance for this choice of trial function, becomes

$$\frac{E_\phi(r)}{r_0^2}=\left\{1+\left[\frac{U_1-U_0}{U_1-U_2}\right]\frac{J_{5/2}(\rho_1\delta)}{J_{9/2}(\rho_1\delta)}\sigma^2\right\}\sigma(1-\sigma^2)^{\frac{1}{2}}. \quad (58)$$

V. THEORETICAL AND EXPERIMENTAL RESULTS

The inductive susceptance of the circular obstacle was calculated for values of δ between 0.5 and 1.

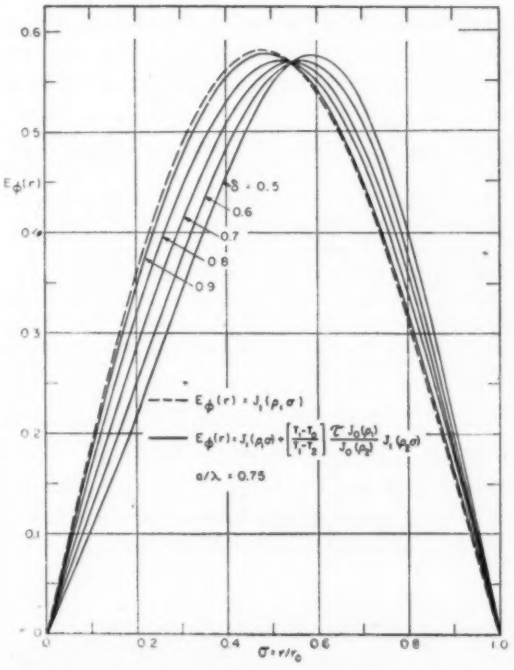


FIG. 7. Assumed electric field in obstacle aperture, first trial function.

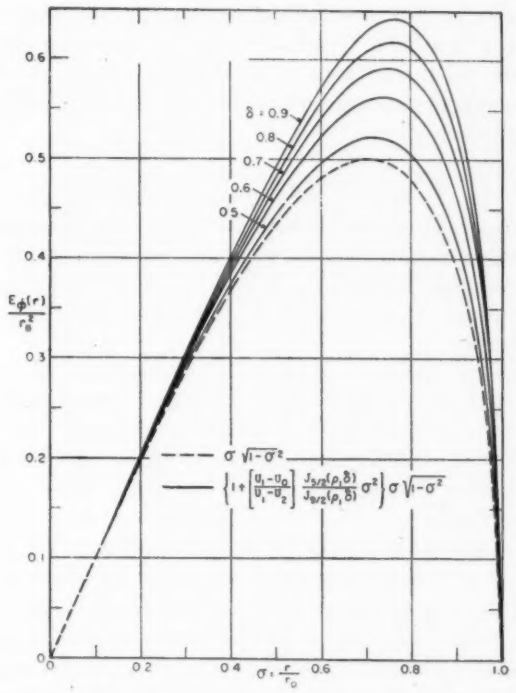


FIG. 8. Assumed electric field in obstacle aperture, second trial function.

Smaller values of δ are of little interest because little energy is transmitted through small apertures. Calculations of the lower bound expression for the susceptance [Eq. (35) as the trial function] and the upper bound expression for the susceptance [Eqs. (46) and (58) used as trial functions] were carried out for $a/\lambda=0.75$. The zero-order and first-order susceptances and also the experimental results are plotted on Figs. 3, 4, and 5. Table I consists of the experimental and theoretical data used to construct these curves. It is observed that the first-order susceptance determined by the lower bound variational formulation and the first-order susceptance derived from the assumed aperture electric field of Eq. (46) in the upper bound formulation are in excellent agreement with the experimental results. The difference between the zero- and the first-order susceptances is a convenient measure of the closeness of the assumed field to the correct field.

It is of particular interest to observe the manner in which a particular trial function varies with the size of the obstacle aperture. In Figs. 6, 7, and 8 the trial functions are plotted as a function of radial position with δ as a parameter. Of course it is impossible to find an exact correlation between the assumed fields and the correct obstacle susceptance because the *actual* aperture field and obstacle current are not known.

The first-order susceptance determined by inserting $E_\phi(r)$ of Eq. (46) in the upper bound variational expression for the obstacle susceptance is plotted in Fig. 9 for values of a/λ equal to 0.70, 0.75, 0.90, 1.0. These theoretical values are compared with the experimental

values and are in excellent agreement for values of δ greater than 0.55. The curves of Fig. 9 can be used to obtain the susceptance of a circular obstacle, provided the obstacle structure meets the requirements set forth by the theory.

VI. EXPERIMENTAL DISCUSSION

The susceptance of the circular obstacle was measured by the resonance-curve method. An excellent description of the experimental technique used for the precise measurement of wave-guide discontinuities is given by Huxley.[22]

The circular wave guide, in which the susceptance of the circular obstacle was measured, consists of brass tubing with 5.750 in. i.d. and 0.125-in. wall. At a frequency of 3000 Mc/sec the wave guide is capable of supporting six propagating modes (TE_{11}, TM_{01}, TE_{21}, TE_{01}, TM_{11}, TE_{31}).

The generator of the "TE_{01} transmission line" consists of a shielded loop of $\frac{3}{16}$ in. diameter, constructed from $\frac{1}{32}$ in. o.d. coaxial line. The generator is mounted on the face of one of the two noncontacting short-circuiting plungers. The loop is oriented to couple maximum energy to the dominant circular-electric mode.

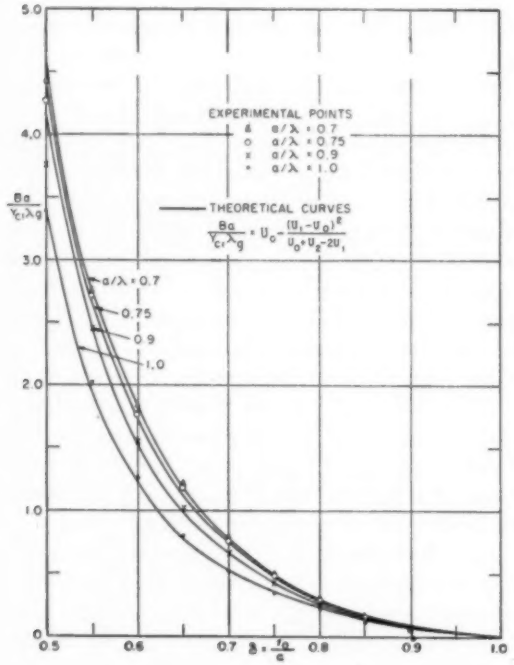


FIG. 9. Obstacle susceptance with a/λ as parameter.

[22] L. G. H. Huxley, *Wave Guides* (Macmillan Company, New York, 1947), pp. 241–246.

TABLE II. Obstacle thickness in terms of guide wavelength.

a/λ	λ cm calculated	λ_g cm measured	t/λ_g
0.70	10.441	21.262	0.00191
0.75	9.737	16.700	0.00243
0.90	8.150	11.110	0.00366
1.00	7.321	9.245	0.00440

The detector consists of four identical shielded loops similar to the excitation loop and uniformly spaced about the circumference of the wave guide. The plane of the detecting loops is oriented in the transverse plane and adjusted so that the currents excited on the loops are codirectional. Individual detector loop outputs of equal phase and amplitude are connected together and to the radiofrequency receiver.

Owing to the discriminating properties of the detector and the damping effect of the lossy material behind each noncontacting short-circuiting plunger, the TE_{01} mode was observed to be approximately 60 db above extraneous modes over a free-space-wavelength range from 7.30 to 11.90 cm.

The circular obstacles were fabricated from brass flat stock of 0.016 in. thickness. The obstacle thickness t in terms of guide wavelength is given in Table II.

It is estimated that the susceptances were measured with an order of accuracy of 0.1 percent.

VII. CONCLUSIONS

Lower bound and upper bound expressions have been obtained for the susceptance of a circular obstacle excited by the TE_{01} mode in circular wave guide. The assumed fields used in the variational formulations are plotted to give a qualitative description of the actual fields that exist on the obstacle. The discrepancy between experimental and theoretical results is probably due to:

1. Approximate nature of the theoretical values.
2. The finite thickness of the obstacle.
3. Finite conductivity of the obstacle.
4. Ellipticity of the circular wave guide.
5. Presence of lower order propagating modes.
6. Difficulty in keeping the thin obstacle planar.
7. Slight asymmetry of the obstacle aperture with respect to the guide axis.
8. Typical errors involved in the measurement of admittance by the resonance-curve method.

ACKNOWLEDGMENT

The author wishes to thank Dr. James E. Storer for his reading and helpful criticism of this paper. The lengthy numerical computations were carried out by Miss Lesley Gerould and Miss Mary Tynan.

Bulk Viscosity of the System Polystyrene—Diethyl Benzene*

F. Bueche
Department of Physics, University of Wyoming, Laramie, Wyoming
(Received September 29, 1952)

The bulk viscosity of the system polystyrene-diethyl benzene has been measured over the concentration range from 8 to 80 weight percent polymer and a temperature range from 30 to 130°C. A single sample of polymer was used. The activation energy for viscous flow is found to vary from about 2 to 100 kcal/mole as one goes from pure solvent to pure polymer. Using a series of 15 fractions of polystyrene ranging from 50 000 to 10^6, the molecular weight dependence for the viscosity of this system was measured at two polymer concentrations, 14 and 44 percent. The results are compared in a qualitative way with what one would expect from existing theory.

INTRODUCTION AND EXPERIMENTAL

THE bulk viscosity of bulk polymers and polymer plasticizer mixtures is of considerable interest for both theoretical and practical reasons. Despite this importance the amount of significant data available is severely limited. Only recently has it been shown that the early data on polyesters[1] are not typical of all polymers[2] as was frequently assumed. There still exists a wide gap in our knowledge on this subject.

Data are now available for the viscous behavior of pure polystyrene and polyisobutylene over large molecular weight and temperature ranges.[2] Some success has been attained in interpreting this data from a molecular standpoint.[3] However, the large region comprising the viscous behavior of plasticized polymer is still essentially unexplored except in a few instances.[4] It is the purpose of this paper to study the effect of temperature and molecular weight on the bulk viscosity of the system polystyrene-diethyl benzene. The viscous behavior of the pure polymer is now fairly well known[2] and the behavior at very low polymer concentrations is quite adequately understood. For this reason we shall confine our work to the region of 8 to 80 weight percent polymer.

The first series of measurements was undertaken to find the effect of plasticizer concentration and temperature on the bulk viscosity of a single sample of polystyrene. We chose a Dow Styron having an $M_v = 540\,000$ and an $M_w = 720\,000$. Its T_g was 84 ± 1°C as determined from volume expansion measurements. The material was washed repeatedly with filtered methanol, dried under vacuum, and placed in test tubes. After adding freshly distilled diethyl benzene, the tubes were sealed off and heated at 90°C until a homogenous mixture was obtained. This process usually took two to three days. No significant degradation was found by running dilute solution viscosities before and after the runs.

A representative portion of the data is shown in Fig. 1 where $\log\eta$ is plotted as a function of $1/T$ where η is the bulk viscosity in poise. The different curves correspond to varying polymer concentrations. Only about one-fourth of the data is shown so as not to complicate unduly the figure. The remainder is similar and of about the same accuracy. Although curvature is indicated when plotted on a larger scale, we have, for ease in graphing the data, used straight lines through the points. Table I gives a numerical listing of some of this same data.

The viscosities were run in straight capillary tubes which had been calibrated from their geometrical dimensions.[5] The more fluid samples were run in viscometers which consisted of bulbs blown in straight capillaries. These were calibrated using glycerine solutions. It is estimated that the data are accurate to about ±5 percent. The chief errors are in the calibration procedure and in the effects of evaporation of plasticizer. Tem-

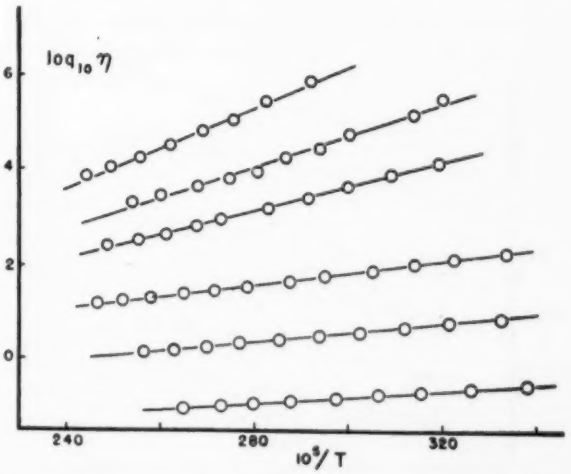


FIG. 1. A representative portion of the viscosity data for a single sample of polystyrene at varying diethyl benzene concentrations, i.e., 80.3, 71.0, 62.0, 38.0, 23.5, and 8.9 weight percent polymer.

* This work was done while the author was employed as a Research Associate at Cornell University working on a contract supported by the synthetic rubber section of the Reconstruction Finance Corporation in connection with the U. S. Government's synthetic rubber program.

[1] P. J. Flory, J. Am. Chem. Soc. **62**, 1057 (1940).
[2] T. G. Fox and P. J. Flory, J. Phys. Coll. Chem. **55**, 221 (1951) and J. Appl. Phys. **21**, 581 (1950).
[3] F. Bueche, J. Chem. Phys. **20**, 1959 (1952).
[4] Bestul, Belcher, Quinn, and Bryant, J. Phys. Chem. **56**, 432 (1952). See also reference 7.
[5] T. G. Fox and P. J. Flory, J. Am. Chem. Soc. **70**, 2384 (1948).

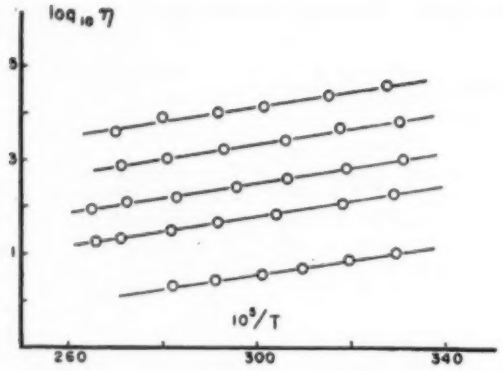


FIG. 2. A representative portion of the viscosity data taken at 44 weight percent polymer with varying molecular weights, i.e., 775, 540, 300, 175, and 52 thousand.

perature control was maintained by immersing the sample and capillary in a thermostated oil bath.

All measurements were taken at very low shear rates. Test runs at varying pressures indicated that increase of flow times by a factor of ten still caused no appreciable change in measured viscosity. It is therefore assumed that the results given here are very close to those one would obtain at zero rate of shear.

A second and third set of measurements was performed in order to determine the effect of molecular weight on bulk viscosity at two different plasticizer concentrations, namely 43.6 and 14.2 percent polymer by weight. The fractions were prepared by a double fractionation from 2-butanone by methanol. The first fractionation was done in three sections with an initial concentration of 1.0 percent. For the second fractionation an initial concentration of 0.7 percent was used.

Molecular weights of the samples were determined from their intrinsic viscosity in benzene at 30°C. The relation used was[6]

$$\log M_v = (\log[\eta] + 4.01/0.73).$$

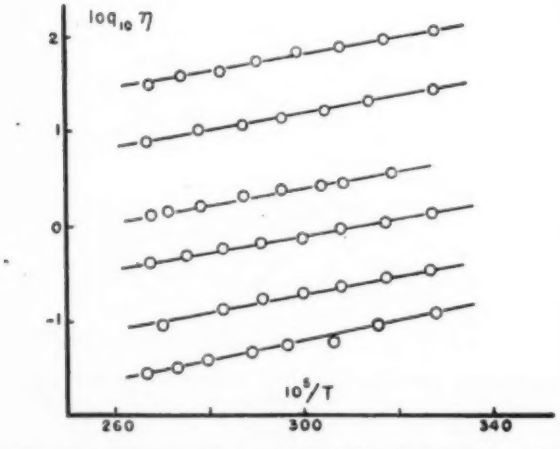


FIG. 3. A representative portion of the viscosity data taken at 14 weight percent polymer with varying molecular weights, i.e., 1170, 870, 615, 262, 143, and 59 thousand.

[6] R. H. Ewart and H. C. Tingey. Abstracts of papers presented at 111th meeting of the Am. Chem. Soc. April 14 (1947).

Some degradation during the mixing process was observed for the highest molecular weights. It was small, however, and the molecular weight used for the figures was found from the dilute solution viscosities after the runs were completed.

The data are presented in Figs. 2 and 3 and Table II. Again we have not included all of the data in order that the figures will not be unduly crowded. However, the data shown are representative. To show the molecular weight dependence more clearly, the data for two temperatures, 30 and 100°C, are plotted in Figs. 4 and 5.

RESULTS AND DISCUSSION

It has been shown[3] that at least to a first approximation one may represent the bulk viscosity of a polymer system by

$$\eta = (R^2/M)(\rho AkT/16a^2P)N^*, \qquad (1)$$

provided the polymer is not too dilute. The symbols are as follows: R^2, mean square end to end distance of the chain; M, molecular weight; ρ, density of polymer in the mixture; A, Avogadro's number; P, segmental jumping frequency; $(3a^2)^{\frac{1}{2}}$, average distance moved by a segment in one jump; and N^* is an effective number of segments per molecule. Its meaning is discussed quite fully in the reference cited.

TABLE I. Bulk viscosity of polystyrene as a function of temperature and diethyl benzene concentration.

	logη at the indicated polymer concentration (g/100 g)										
T°C	8.9	13.3	23.5	29.1	38.0	45.5	53.5	62.0	71.0	75.5	80.3
35.0	−0.63	−0.24	0.78	1.30	2.14	2.73	3.58	4.32	5.69	6.55	···
50.0	−0.73	−0.33	0.64	1.10	1.92	2.49	3.25	3.90	5.06	5.84	6.85
70.0	−0.87	−0.48	0.45	0.92	1.68	2.22	2.88	3.38	4.33	4.96	5.82
90.0	−0.98	−0.58	0.30	0.78	1.49	2.04	2.59	2.97	3.83	4.35	5.04
110.0	−1.08	−0.68	0.16	0.65	1.32	1.90	2.36	2.63	3.45	3.93	4.46

Although ρ is the only quantity in the above relation which depends directly on the concentration of polymer, one would expect a rather strong dependence of P on diluent concentration. This is confirmed by many experiments on dielectric resonance phenomena as well as by less conclusive arguments based on the lowering of T_g by added diluent. Also, it will be shown that N^* is concentration dependent.

The data showing the dependence of viscosity on polymer concentration are best illustrated by a plot of $\log_{10}\eta$ *versus* C as has been done in Fig. 6. The two curves represent data taken at 39 and 97°C. One of the most striking features about these curves is that they show a pronounced leveling off at intermediate concentrations and then a rapid rise at higher concentrations. In fact, at the concentrations indicated by arrows, the viscosity should be of the order of 10^{12}. This is inferred from published data on the viscosity of polystyrene and from other data showing how much T_g is

TABLE II. Bulk viscosity of polystyrene-diethyl benzene solutions at 30°C and 14 and 44 percent polymer.

$M/10^4$	$\log\eta$ 14%	$\log\eta$ 44%
52		1.02
59	−0.93	
77		1.38
115	−0.70	1.65
143	−0.42	1.94
175	−0.26	2.25
212	−0.18	2.57
262	0.22	2.84
300		2.97
310		3.10
345	0.28	
403	0.26	3.50
460		3.72
570	0.52	
615	0.63	4.07
775		4.58
800	1.25	
870	1.45	
1170	2.09	

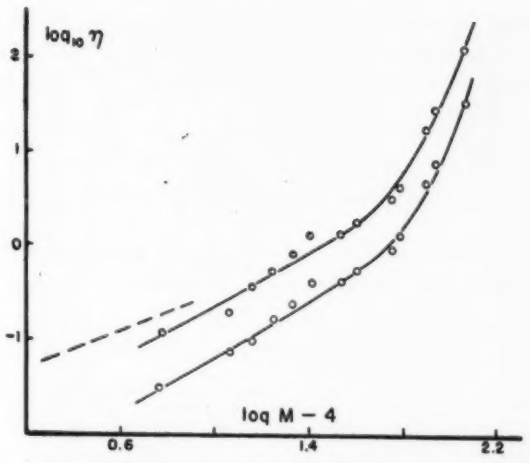


FIG. 5. The variation of bulk viscosity with molecular weight at 14 weight percent polystyrene.

lowered by known amounts of added diluent.[7,2] This behavior has been reported previously.[8] The other end of the plot is established by the fact that pure diethyl benzene has a viscosity of 0.0076 poise at 30°C. Its activation energy for flow is 2.0 kcal/mole.

It is noteworthy that the sharp rise in $\log\eta$ at high concentrations is not observed for rubber-like polymers.[4] Such a situation should not be entirely unexpected since the figure shows that the flat region is longer for higher temperatures. In fact, if the temperature of the polystyrene mixtures were raised as high above the transition temperature T_g as the rubber system was above its T_g, it is quite possible that the flat region would continue to 100 percent polymer. This is more clearly shown by a plot in three dimensions such as is shown in Fig. 7. The data have been extended to 100 percent polymer by making use of the results of Fox *et al.* It is quite obvious from this figure that as one goes to higher temperatures and moves farther from the transition region (represented by the more or less conical tower of Fig. 7) the viscosity-concentration curve flattens out rather rapidly. Because of this, we suspect that the sharp rise at high polymer concentrations is due to a rapid decrease in P near the transition region.

It is often customary to express P in the general form $P=P_0\exp(-E/kT)$ where E is an activation energy. To obtain E from Eq. (1) we need only compute the derivative of $\ln\eta$ with respect to $1/T$ provided we ignore the slow variation arising from the factor T. The experimental data have been treated in this fashion and the result is shown in Fig. 8. The curve represents k times the graphical derivative taken at 50°C. At the concentration indicated by the arrow the activation energy should be about 100 kcal/mole. This follows from the fact that E at T_g is about 100 kcal/mole.[2,7]

We see that a rather sudden rise in E occurs about where the viscosity shows a rapid rise. It would, therefore, seem that the major cause of the viscosity rise is

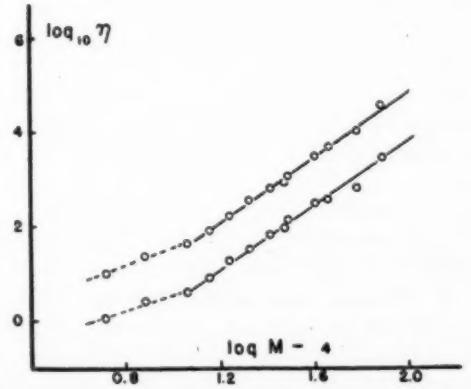


FIG. 4. The variation of bulk viscosity with molecular weight at 44 weight percent polystyrene.

[7] Fox, Schultz, Schultz, and Flory (to be published). I am indebted to these authors for making their data known to me.
[8] T. G. Fox, Phys. Rev. 86, 652 (1952).

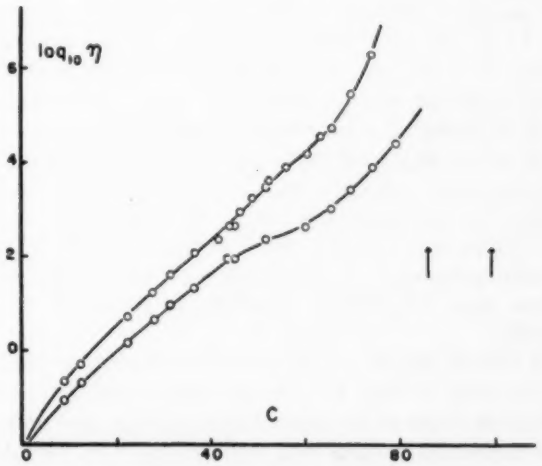


FIG. 6. The bulk viscosity of a single sample of polystyrene in diethyl benzene as a function of weight percent polystyrene.

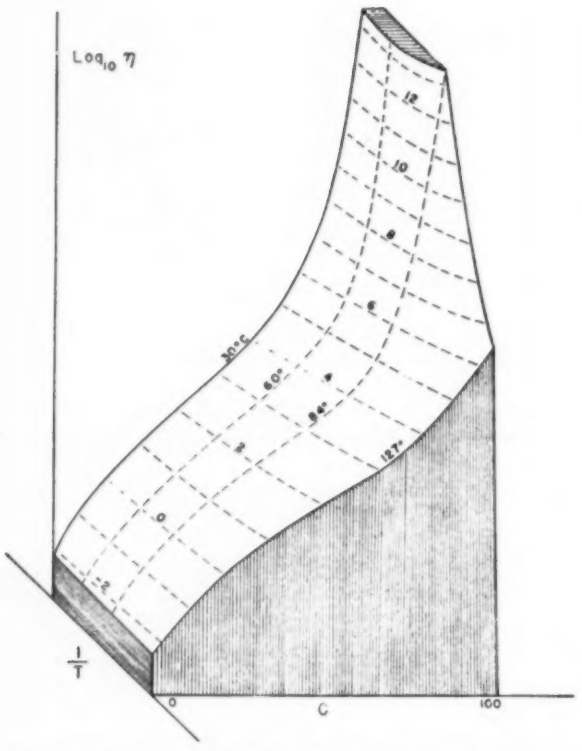


FIG. 7. The variation of bulk viscosity of a single sample of polystyrene with respect to temperature and weight percent polymer. The plot is isometric in the sense that a distance along an axis is proportional to the variable under consideration.

the increase in E. Since E varies less rapidly in regions far removed from T_g, it seems reasonable that the bulk viscosity of rubber-like materials far removed from T_g should not show the sudden rise in viscosity observed here. As we have mentioned earlier, such is actually found to be the case.

The effect of molecular weight on the bulk viscosity is shown in Figs. 4 and 5. At 44 weight percent polymer, Fig. 4, the high molecular weight region follows the relation $\eta \sim M^{3.4}$, which is identical to the result found for the pure polymer by Fox and Flory.[2] For the case $C=14$ percent, the relation is seen not to hold any longer. In this case, the viscosity of the high molecular weight portion actually rises with even a higher power of M. Although it is impossible to assign an exact power of M owing to the curvature, it is of the order of 4.5.

Even more striking is the fact that the break in the curve does not always occur at the same value of M. For the pure polymer it came at about 50 000.[2] At 44 percent polymer it occurs at $M=10^5$ and in the most dilute case, 14 percent polymer, it comes at about 4×10^5.

It should also be pointed out that the lower portion of the data in Fig. 5 is nearly represented by $\eta \sim M$. The broken line in the figure has that slope. Not enough low molecular weight data are available in Fig. 4 to assign an exponent in that case.

These two experimental facts, the shift in the position of the break and the changes in slope with concentration, may be fairly easily understood from a consideration of the theory previously proposed.[3] It was pointed out there that the position of the break occurs at the molecular weight for which, on the average, one polymer molecule has one other molecule so entangled with it that the two must move essentially as one. The precise interpretation is somewhat more complicated but the reader is referred to the original article for a more extended treatment.

With this meaning of the position of the break in mind, it is clear that as the solution becomes more dilute in polymer, the average number of molecules coupled to any one molecule must decrease and the break must, therefore, move to higher molecular weights. In the extreme limit of very dilute solutions, the break will not occur at all since the molecules will be widely separated. This fits in nicely with what our experiments have shown.

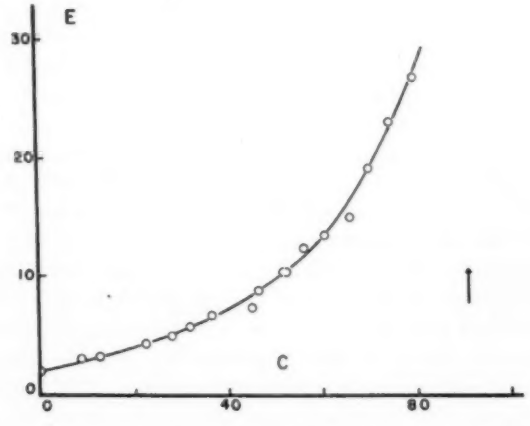


FIG. 8. Apparent energy of activation for viscous flow of a single sample of polystyrene in diethyl benzene as a function of weight percent polymer.

The slope of the curves has been shown to be the result of two factors.[3] Ordinarily one would find that at low molecular weights the slope would be unity, i.e., $\eta \sim M$. When the position of the break is reached one finds that the slope increases to a higher value determined essentially by a slippage factor for the coupled chains. If no slippage occurred the slope would become infinite.

However, another factor must be considered. As the number of chain ends increases (i.e., M getting smaller) the segments of the chains will find movement easier and so P will also vary with molecular weight. This effect is difficult to assess accurately but qualitatively it must increase the slope of the low molecular weight portion and decrease the higher slope. But when the amount of diluent becomes large, so that P will be relatively insensitive to the presence of chain ends, we should expect the lower portion of the curves to take on a slope of unity. An accompanying increase in slope for

the upper portion should occur. This is actually seen to be at least qualitatively correct.

It should be pointed out that this paper has done no more than outline the over-all behavior of the system polystyrene-diethyl benzene. Much more work remains to be done. In particular the molecular weight dependence at various plasticizer concentrations should be extended over a wider range of molecular weights. This should lead to a more coherent picture of the effect of entanglements. Moreover, the molecular weight dependence should also be investigated at a large number of polymer concentrations so as to obtain a clearer idea as to the meaning of the variation of the break in the curves of Figs. 5 and 6. Studies over a wider temperature range would also be of value. However, a different plasticizer should be used for temperatures higher than reported here so as to minimize evaporation effects.

The author is deeply indebted to Mrs. Gloria Potter for her aid in taking the measurements reported here.

JOURNAL OF APPLIED PHYSICS VOLUME 24, NUMBER 4 APRIL, 1953

Positive and Negative Ions in Cathode-Ray Tubes*

C. H. BACHMAN, G. L. HALL,† AND P. A. SILBERG
Physics Department, Syracuse University, Syracuse, New York

(Received December 8, 1952)

The positive and negative ions accompanying the operation of a cathode-ray tube contribute in a number of ways to tube failure. These ions have been studied in standard sealed-off tubes. The identities of the most persistent ions have been determined and ion currents measured. Some attempts to limit or control the ions at their sources have been made.

The greatest offender is hydrogen which is present as both negative and positive ions. It is shown that aluminized screens offer little protection to the ion bombarded phosphor. Uncertainty in the literature as to whether negative m/e, 26 is C_2H_2 or CN seems resolved in favor of C_2H_2.

A NUMBER of deleterious cathode-ray tube phenomena may be traced directly to positive and negative ions active during the tube operation. The identities of the negatives have been ascertained, and a number of possible ion forming mechanisms and negative ion sources have been postulated.[1–5] A recent paper[6] describes a hollow cathode type of apparatus designed to obtain information concerning the positive ions. The present paper gives further results of this study of background ions.

I. DOUBLE-ENDED, HOLLOW-CATHODE DEMOUNTABLE APPARATUS

With this equipment, reproduced again in Fig. 1, it was determined that the greater part of the positive ions are formed at or near the first positive electrode passed by the electrons. Furthermore, these positives arrive at the cathode having very closely the energy given by the potential with respect to the cathode of that electrode. (This infers a restricted initial energy spread). Post-mortem examinations[1] of cathode-ray tubes have shown apertures etched by what must have been positive ions from the bulb or screen. However, the present retarding potential measurements show that such ion current must be very small compared to that arising at the first positive aperture, which in this type of gun draws the larger part of the electron emission.

With the ion beam focused at the screen and deflected by a transverse magnetic field, the positive ion components through the hollow cathode may be separated according to their values of m/e. The most intense ions were found to be hydrogen, m/e 1 and 2. The remainder of the spectrum was variable. Sometimes as many as eight components were visible in fluorescence or observed by screen burning. The number of components and the sharpness of spot were quite sensitive to pressure in the system, and well-focused spectral spots were the exception. The usual type of spot distortion

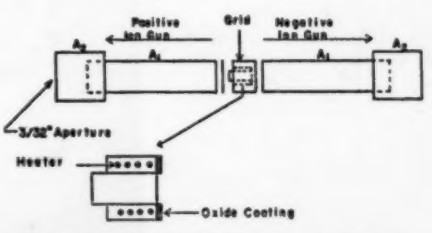


FIG. 1. Double-ended hollow cathode gun.

* Portions of this work have been sponsored by the General Electric Company and by the Office of Ordnance Research, U. S. Army.

† Now at Federal Telecommunications Laboratory, Nutley, New Jersey.

[1] C. H. Bachman and C. W. Carnahan, Proc. Inst. Radio Engrs. 26, 529 (1938).

[2] L. F. Broadway and A. F. Pearce, Proc. Phys. Soc. (London) 51, 335 (1939).

[3] C. H. Bachman, J. Appl. Phys. 11, 83 (1940).

[4] H. Schaefer and W. Walcher, Z. Physik 121, 679 (1943).

[5] R. H. Sloane and C. S. Watt, Proc. Phys. Soc. (London) 61, 217 (1948).

[6] Bachman, Eubank, and Hall. J. Appl. Phys. 22, 1208 (1951).

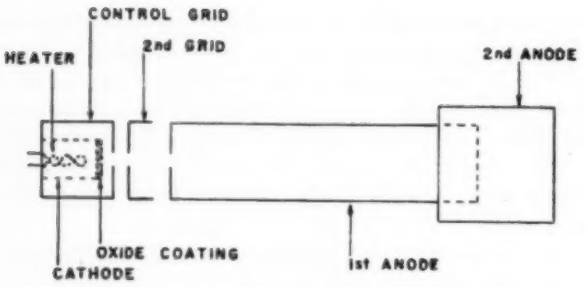


FIG. 2. Standard electrostatically focused electron gun.

consisted of elongation in the line of the spectrum resulting from ion formation along the beam.

In one experiment the positive ions were caused to strike the masking aperture disk on the positive ion gun, and the potentials of the other focusing electrodes were run positive so as to focus negative ions. Two ion components were observed definitely which only could have been negatives formed at the aperture by the positive ion bombardment. Control was not good and the identities of these two were not determined.

Variations in the relative intensities of the ions could be obtained by either sparking or by heating the glass walls of the system. Such procedure did not seem capable of giving reliable information as to the gas driven from the walls. After prolonged sparking or heating, gas evolution diminished although the area became "rejuvenated" after a rest period of many minutes or hours.

A similar effect occurred at the aperture being bombarded by electrons. The ion yield diminished noticeably after several hours of operation but appeared with full vigor after a night's rest (see Sec. III).

Attempts to isolate the original source of the ions included building three complete glass vacuum systems for operation with octoil, silicone oil, and mercury. Liquid air and CO_2 were checked as refrigerants. No significant differences were found except with the Hg system. Here CO_2 refrigerant had no effect while liquid air on the trap seemed to offer a definite reduction in ion yields. Baking of the apparatus was, of course,

TABLE I.

Positives		Negatives	
m/e	Identity	m/e	Identity
1*	H_1^+	1*	H_1^-
2*	H_2^+		
3*	H_3^+		
4	C^{+++}		
6	C^{++}		
7	N^{++}		
8	O^{++}		
12*	C^+	12*	C^-
14	N^+ or CO^{++}		
16*	O^+	16*	O^-
28*	CO^+	26*	$C_2H_2^-$
		32	O_2^- or S^-
		35–37	Cl

(Other heavier ions apparently complex and not consistent from tube to tube.)

impossible. Gas pressures were in the range of 10^{-4} mm Hg.

Fluorescent screens for positive ion reception were originally placed on metal to eliminate charging problems which might arise with glass. Tests of phosphor on glass, however, showed that although such screens were often charge sensitive, they usually received the positive ion spectra with little trouble.

The information obtained with this demountable equipment pointed to the possibility of continuing the work with sealed-off cathode-ray tubes produced according to standard procedure. From the practical standpoint this is to be desired. Results from a demountable equipment which could not be outgassed properly would always be influenced by the contaminants of that system. Results from sealed-off production tubes, on the other hand, would reflect the ion characteristics which were actually of interest.

II. ION IDENTITIES IN SEALED-OFF TUBES

The gun structure used in this work is shown in Fig. 2. It is an electrostatically focused gun consisting of cathode and control grid, accelerating aperature and first and second anodes. The accelerating aperture normally operates at a few hundred volts, the first and

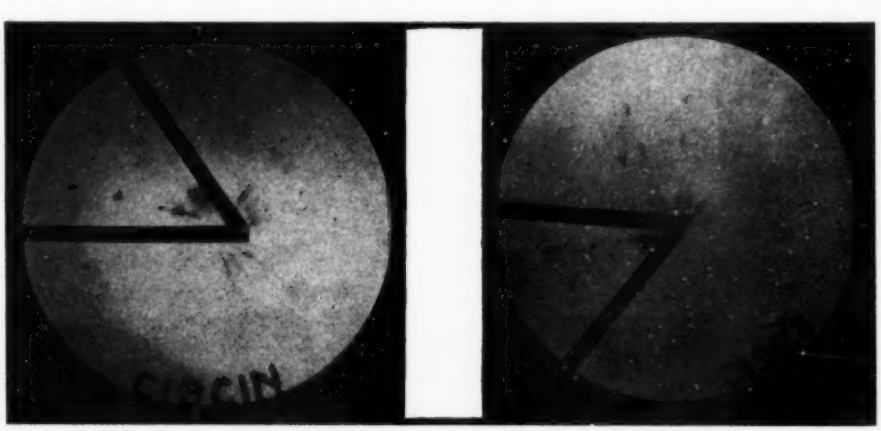


FIG. 3. Displays of adjacent positive and negative ion burn spectra.

second anodes much higher, say 2000 and 8000 volts, respectively, for focus.

The cathode, control grid, and accelerating electrode serve to generate positive ions in the manner demonstrated in the demountable apparatus. If the first and second anodes are supplied with negative potentials, positive ions will be drawn from the accelerating electrode and focused on the screen. Here they may be inspected in detail by magnetic deflection. With normal positive potentials, negative ions may be studied. It is possible to place side by side the ion burn spectra of both the negative and positives of a given tube (Fig. 3). In the figures the pi-sections contain such pairs as spectra. The more complex spectrum is that of the positives, and the magnet has been adjusted so that the outermost spots (H_1^+ and H_1^-) are about equally deflected.

A large number of such tubes have been examined, Table I lists the mass numbers observed. Those values starred are most commonly found, and may be considered typical of average tubes. It is interesting that

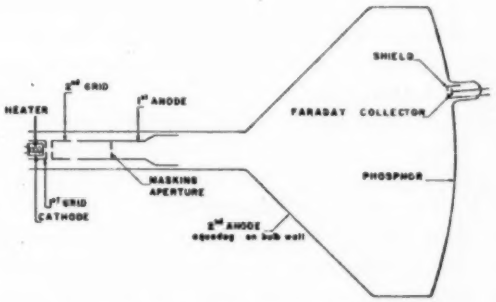


FIG. 4. Sealed-off tube with Faraday cage.

some of the persistent negative ions of early day tubes —chlorine and m/e 32, (O_2 or S) for examples, seem to have been nearly eliminated.

III. ION CURRENTS IN SEALED-OFF TUBES

Ion currents were measured in special tubes with Faraday cage collectors sealed into the tube faces in an off-center position as shown in Fig. 4. The magnet current was adjusted to bring the desired ion component onto the collector slit, and the currents were measured with a dc amplifier. Typical curves for positive and negative ions are shown in Figs. 5 and 6. The appearance of m/e 0.75 is due to ring structure and will be discussed later. Ion yields as functions of various tube operating potentials have been taken and will be the subject of another paper.

Some observations on rejuvenation of ion sources were mentioned in Sec. I. This action was seen more clearly in connection with the Faraday cage tubes. Measurements of ion currents with such a tube entail its operation for long periods under exactly the same electrical conditions and placement in the magnet.

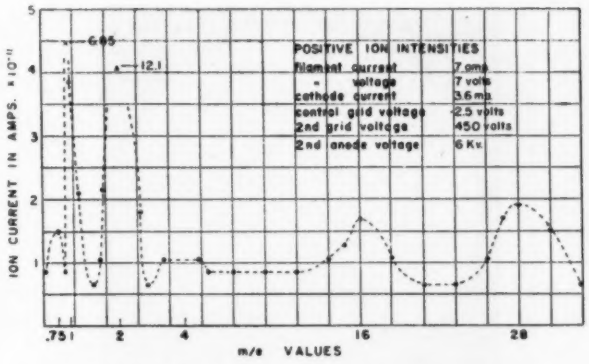


FIG. 5. Typical positive ion intensity curve.

During such time the same portion of the aperture is being bombarded with electrons.

While making such measurements it was noted that the ion intensities decreased with time; the recovery by an overnight rest period was only partial. It was discovered that the entire spectrum could be rejuvenated by a few minutes operation with the electron beam accelerated to the fluorescent screen.

Apparently electron bombardment of the remaining gun apertures, Faraday cage, and phosphor, released gas some of which found its way to the accelerating electrode restoring it effectively as a positive ion emitter. This rejuvenation was quicker and much more effective than that resulting from a tube rest of many hours.

In addition to the gradual decrease of ion intensities with time, a short time effect has been noted on a number of tubes. If all gun potentials are present prior to electron bombardment of the aperture, the ion spectrum which appears in the first few seconds is extremely complex and intense. Before the components can even be counted, the intensities drop and the spectrum simplifies to the usual pattern.

The degree of vacuum of cathode-ray tubes is usually measured in terms of the "gas ratio"; the ratio of ion to electron current when the cathode and nearby electrodes are arranged to function as an ionization gauge. Tubes in this investigation commonly had gas

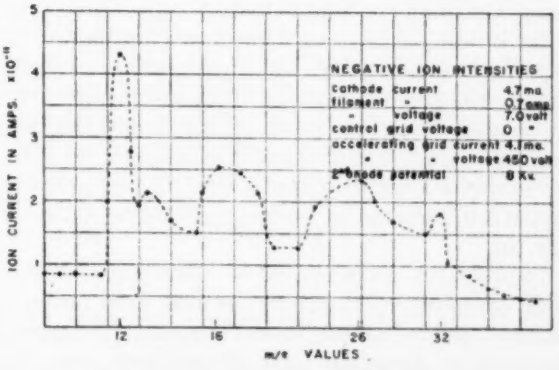


FIG. 6. Typical negative ion intensity curve.

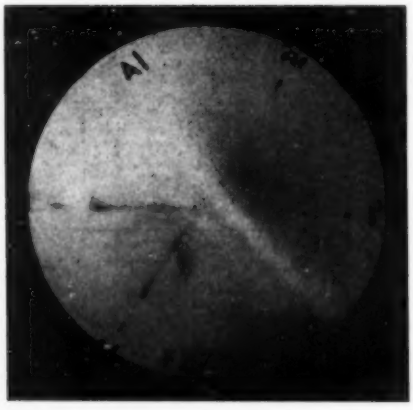


FIG. 7. Ion burns on tube having one-half the screen area aluminized.

ratios from 10^{-4} to 10^{-7}. (See discussion in Sec. VII for the significance of this measurement.)

IV. INFORMATION FROM SPECIAL TUBES

Many special tubes and treatments were investigated in trying to determine and eliminate the source of ions. The results of some of these tests are described here.

A tube was built having no cathode, heater, or control grid. An rf induction coil external to the tube was arranged to heat the accelerating electrode. Focus voltages and magnet were applied and the aperture heated by induction. This technique was difficult since the rf heater developed heat, outgassed the aperture, and caused electrical gas discharge all at the same time. Observation periods were confined to a fraction of a second after turning off the heater. A number of positive ions were observed in fluorescence; H_1 and H_2 were the only ones identified. These ions gave ring images of the accelerating aperture. Such rings have been mentioned in earlier literature and will be discussed later in this paper. The technique did not lend itself to a search for negative ions.

Tests were made of ion content of tubes having different interior surfaces. There were no significant differences between graphite-coated iron and glass tubes. Graphite coatings definitely contributed greatly to the carbon ion content. Tubes coated with aluminum or silver paste showed either no carbon or greatly minimized carbon intensities. This was further verified by increased carbon positive ion intensities which resulted when a spark coil was held up to the tube at the graphite edge near the screen. It was also noted that when there is no carbon there is no component m/e, 26.

Comparisons were made of tungsten filament *vs* oxide cathodes, hydrogen *vs* vacuum fired parts, and mercury pump *vs* oil pump exhaust. Accelerating apertures of zirconium and of tantalum were tried. None of these tests gave significant differences. It must be remembered that these comparisons were necessarily qualitative.

It is of interest that in one tungsten filament tube the filament had been helium fired to set it during assembly. This tube produced a strong helium ion even after many hours of operation.

Several tubes were built in which only half of the screen and bulb were aluminized. The decrease in ion burning of the aluminized portion is shown in Fig. 7. Referring to diametral line 2, for example, the rather complete set of spots was burned into the unaluminized portion of the screen. The magnet was then reversed, and the exposure time increased until the H_1 burns (outermost spot) were comparable.

The positives H_1 and H_2 were visible in fluorescence through the aluminum down to about 2.5 kv. In addition to requiring about 6 times the exposure time for a comparable hydrogen burn, the heavier ions burns are eliminated by the aluminum, at least for exposure times normal for hydrogen. Note that whereas H_2 gives a more intense burn in the normal screen, the aluminum film retards H_2 allowing H_1 to give the greater burn. This effect is much greater than can be explained by the known thickness variations of the aluminum.

Although negative H could be observed through the aluminized area, the intensity was too low to burn through the aluminum with exposures up to several hours. There is a question as to whether the ion burn through the aluminum results from true penetration of the ion or transfer of its energy through the aluminum to the phosphor. All the ion components have the same energy, of course. Ion currents and corresponding burns seem closely related. It would seem then from Fig. 7, that penetration is involved since in the case of simple energy transfer all ions would be expected to cause roughly equal burning to the extent of their current densities.

In some tubes, it was noticed occasionally that some heavy ions would appear only under very high voltage operation, say at 10 kv, and that these would disappear at lower voltages. This effect appeared to be associated with increased leakage between electrodes either along the beads or through cold emission from electrode edges. (Note the paragraph above describing increased carbon ion for sparking of the graphite edge.)

A tube was built in which the usual cup-shaped accelerating electrode was replaced with two aperture

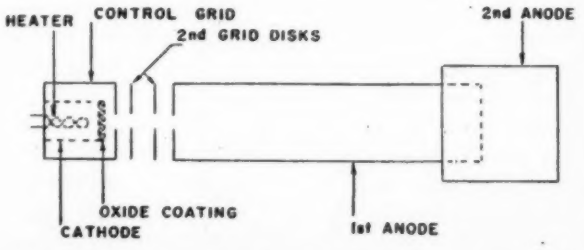


FIG. 8. Electron gun for observing ion and electron paths.

disks as shown in Fig. 8. Small amounts of phosphor were placed over parts of the disks for the purpose of observing the areas bombarded. This tube eventually developed sufficient gas that the paths of both electrons and positive ions could be observed as gas beams and was valuable in indicating the paths of particles involved in positive ion beam formation.

In general, experiments with the sealed-off tubes showed that whereas the negative ion components focused well and were consistently placed for measurement, the positives arriving at the screen often appeared to be confined to a region rather than a spot, usually focused to a radial line together with a spot and often were displaced laterally from the radial spectrum direction. These effects will be discussed later.

V. NEGATIVE IONS

The most common negative ions are H, C, O, and C_2H_2. These appear in tubes having tungsten cathodes as well as oxide cathode types. Negatives are in general sharply focused (compared to the positives) indicating a sharply determined source area and little energy spread (i.e., well under 100 volts). The sources are the cathode and in some cases apertures. In the current set of experiments the only ion contributing aperture which could have been focused was the accelerating aperture. No negatives from this aperture were observed.

Contrast this with Fig. 9 showing spectra of a tungsten filament tube reproduced from an earlier paper.[1] The method of formation of the spot and ring structure has been explained[3] chiefly on the basis of negatives arising from positive ion bombardment. In that tube there was no accelerating aperture, and the aperture serving as the negative ion source was shown to be the control grid. Table II gives m/e values. (With the deflecting field used, H_1 would have been deflected out of the picture.) The heavier values are of questionable accuracy since these measurements made recently were necessarily carried out on a small photograph. Details of the structure of the components are given. Only the three lightest, 12, 16, and 24 show center spots as evidence of coming from the filament as well as the aperture.

Negative ions of both m/e, 16 and m/e, 32 have been reported for oxide cathode gun structures under conditions in which high energy bombarding positives were present. Sloane and Watt[5] find both ions from the cathode as well as the first accelerating mesh of their mass spectrometer. Plumlee and Smith,[7] in what seem to be experimental conditions similar to those of Sloane and Watt, fail to find either ion. They use this fact together with the dependence of O_2, 32^+ yield upon electron emission, as evidence for the emission of neutral oxygen molecules from the cathode.

TABLE II.

m/e	Intensity	
12	weak	(ring) (center dot) and (side dot)
16	strong	(ring) (center dot) and (side dot)
24	weak	(ring) (center dot) and (side dot)
26	strong	(ring) (no center dot) (side dot)
32	weak	(concentric rings) (no center dot) (side dot)
36	strong	(diffuse ring) (no center dot) (no side dot)
44+	weak	(concentric rings) (no center dot) (no side dot)
63	weak	(ring) (no center dot) (side dot)

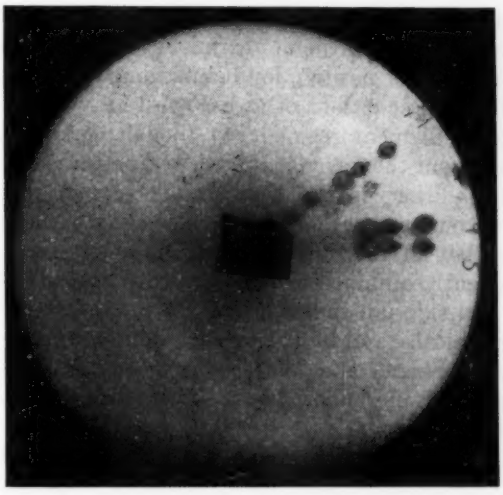


FIG. 9. Ring spectra of tungsten filament tube.

Although m/e, 16 from the tungsten filament is present here as shown by the center spot, m/e, 32 appears to come only from the aperture (no center spot). If O_2 is emitted by the filament, it must be dissociated immediately. Ions m/e, 26, 44, and 63 likewise come only from the aperture, having no center dot, while the unidentified m/e 36 has neither side dot nor center dot.

That bombarding positives were present in this earlier tube is shown by the side spots in some rings, but as Sloane and Watt have pointed out, dependence upon bombarding positives for those rings *without* spots would require some selectivity not yet shown between incident positives and emergent negatives. One explanation may be as follows.

The aperture yielding the spectra of Fig. 9 was the control grid, closely spaced to a tungsten filament—its temperature must have been several hundred degrees Centigrade during operation. In the current work the accelerating aperture would have to serve as the corresponding negative ion source. This aperture is separated from the relatively low temperature oxide cathode by the control grid and its temperature must be comparatively low. No negatives, either rings or spots have been observed from this colder aperture.

Comparing Zn and ZnO surfaces, Dart[8] has shown that thermionic positive ion emission (m/e 64) from the Zn is unaffected by electron bombardment, while there is definite increase in ion yield from the ZnO

[7] R. H. Plumlee and L. P. Smith, J. Appl. Phys. **21**, 811 (1950).
[8] F. E. Dart, Phys. Rev. **78**, 761 (1950).

under such bombardment. Our results indicate a similar selective enhancement of thermionically emitted negative ions under positive ion bombardment.

The absence of ions of m/e 26 and heavier from the tungsten filament region even though these ions appeared from the lower temperature aperture, indicates that conditions in the filament region were not conducive to the formation or to the longevity of these ions. Note that in the current work, however, m/e 26 is consistently obtained from oxide coated cathodes. The ring and side dot indicate this ion to be formed both thermionically and by positive ion bombardment. This confirms Sloane and Watt who find a small excess energy tail superimposed on a Maxwellian distribution for this component.

In discussing the diffuseness and broad character of the rings and the degree of focus of some of the spectral spots Sloane and Watt[5] infer that these must be associated with excess initial energies of the ions. Certainly the initial energy and also the direction of emission of the ion plays a part in determining the effective size of the source to be imaged on the screen. If the energy spread is large, however, there will be an appreciable spot distortion in line with magnetic deflection. Such an energy spread would be characterized in these spectra as a *radial* (with respect to tube center) spread, whereas the diffuseness under consideration is symmetrical about the ring center.

The resolution of our spectra is such that in the region m/e 10, a ΔV of 50 volts would have been easily discernable unless it were a tail of very low intensity. The ring thicknesses even if they were radial instead of symmetrical, would correspond to ΔVs of several hundred volts. Sloane and Watt's energy measurements indicate energies well below such values. It would seem that the density distribution in ring structure cannot be explained by initial energy distribution but rather is concerned with the geometry and emitting characteristics of the aperture surface.

One question concerning negatives is that of the identity of m/e 26, attributed to either CN or C_2H_2.[1–5] This is one of the consistent ions. Only rarely does a tube show nitrogen in the positive spectrum whereas carbon and hydrogen are consistent and strong as both negative and positive ions. This would seem evidence favoring C_2H_2 over CN for this mass number.

VI. POSITIVE IONS

The positive ions most commonly found are H_1, H_2, C, O, and CO. These are generated mostly by bombardment of surfaces by electrons. No positive ions were detected from oxide cathodes (the test with the inductively heated sperture would seem to indicate thermionic emission but this is not certain since the focusing voltages could have produced a gas discharge of short duration with accompanying bombardment of the aperture). The electrodes receiving most electrons yield most ions, since electrode potentials are above the value where ion yield varies with potential. This point must be kept in mind when guns of different design are under consideration.

Some of the positives formed by electron bombardment emerge in the direction of the incident electrons and being of opposite charge they retrace the electron path back to the cathode arriving with energy very nearly that given by the potential between the bombarded electrode and the cathode. Estimates of initial energy based upon sharpness of lateral focus (see discussion in previous section) indicate initial velocities to be certainly not greater than 100 volts.

Ions formed in bulb space if initially directed into the tube neck may join the axial beam and be accelerated toward the cathode until intercepted by apertures or the cathode itself. Bulb space positives may also be attracted to the fluorescent screen when it is running below anode potential as a result of insufficient secondary electron emission (condition for sticking potential). This may give rise to large burn patterns in the shape of arcs and X's.

Although not pertinent to positive ion behavior in normally operating tubes a number of pecularities were encountered in studying positives in sealed-off tubes.

The ions, though well-focused laterally, tend to smear radially along the spectrum. The extent of this elongation is too great to be accounted for by energy distribution as pointed out earlier and, therefore, must be peculiar to the source geometry. The intensity measurements made with the Faraday cage tube occasionally indicated "extra" unaccountable spots—for example, under one set of conditions a spot appeared at m/e of about 0.75. Occasionally a spot would be slightly displaced laterally from normal spectrum "line up." Sometimes a cluster of spots appeared. Finally, although given m/e spots generally dropped into place well, there was a tendency to "drift within small limits."

That the explanation for all these phenomena lies in the ring nature of the source was nicely confirmed by observation of the tube of Fig. 8. This ring nature is well shown in Fig. 9 for the case of negative ions resulting from positive ion bombardment of an aperture. The appearance of rings of positive ions from the inductively heated aperture was mentioned.

The tube of Fig. 8 was of open structure and became just gassy enough that electron and ion paths could be observed directly. For positive ion analysis the anodes operate with negative potentials. The electrons pass through the various aperture holes from the cathode until faced with the repelling field of the first anode aperture. They are then turned backward in mirror action thus bombarding the "screen side" of the last aperture. This bombardment causes positive ion emission, the positives being drawn forward and focused upon the screen.

The magnetic field used to analyze these ions penetrates to the "source region." This penetration is not

sufficient to affect the ions to any degree in this region but it does influence the electron beam undergoing the mirror action. The electron beam normally would be turned back uniformly in all radial directions. The result of the superimposed field, however, is to create a "one wayness" to the electrons and thus the emitted ions. This "one wayness" appears as a radial extension in the burn spectrum.

Under most operating conditions a focused spot of electrons appears on the bombarded aperture and may be seen in fluorescence on a screened area of the aperture, or under sufficient bombardment may be seen in incandescence. In either case positive ion emission is greatest from this region, and the usual radially smeared image shows a concentrated spot area. This spot, of course, is sensitive to the magnetic field strength and thus for different operating conditions the observed ion spot "shifts" following the shifting of the source. The erratic lateral spot displacement and clusters of spots can be accounted for as preferred ion emission from different regions on the aperture disk. The double images (as the m/e, 1 and 0.75) are believed to be indications of the same m/e from opposite areas of the disk. Under these conditions other ions would not give "ghosts" having these same ratios of course, since this is a geometrical displacement.

VII. DISCUSSION

When the combined negative ion—positive ion situation is reviewed, certain conclusions are possible. In this over-all picture we find the persistent ions to be H_1^{+-}, H_2^{+}, C^{+-}, O^{+-}, $C_2H_2^{-}$, CO^{+}. By the nature of our investigation it was necessary for these materials to appear on (or very near) the surface of the gun apertures. Although there is evidence that the supply diminishes under electron bombardment, it is in some manner replenished after a suitable rest period or when other areas evolve their gas under electron bombardment.

That the ions can be accounted for by materials diffusing out of the metal of the aperture seems doubtful in the face of evidence such as the test of vacuum *vs* hydrogen firing, and the use of several aperture materials. It seems certain that the migration occurs through space. This does not preclude surface migration and diffusion from the solid material but no evidence points conclusively to these factors whereas a good example of space migration is the carbon which seems to originate at graphite edges. Our results, then, are representative of occluded tube gases rather than of the bombarded aperture, although some of the less consistent spots may be peculiar to the aperture.

A rather violent degree of ionization is indicated by the lack of complex m/e values, the concentration of carbon and hydrogen, and the multiple ionization observed. These facts might also tend to indicate that hydrocarbons were originally present. Plumlee and Smith[8] point out the contributions of oxide cathodes to such contamination. However, the inductively heated cathodeless tube experiment eliminates the heater and oxide coating as a hydrocarbon source, and the oil *vs* mercury pump experiments, although less positive evidence would seem to eliminate vapor backstreaming as a potential hydrocarbon source. One unexplored possibility is the gas atmosphere surrounding sealing and exhaust equipment as a result of open gas flames.

However, the nearly complete control of carbon by elimination of graphite bulb coating does not quite fit the picture of hydrocarbons as the primary ion source. The feather edges of graphite coats are more prone to high field destruction, and disintegration of such an edge would be a way in which the graphite could yield its carbon contribution. In any event some other explanation must be found for the other ions, particularly the intense hydrogen, of graphiteless tubes.

Another possible source of the ions lies in the residual gases diffusing from the various metal and glass surfaces not adequately outgassed in processing. These would include bulb walls, beads, side rods, headers, and various parts not well heated during induction firing or which run hot during operation.

The plentiful diffusion of H_2O and CO_2 from glass and CO, H_2, N_2, and O_2 from metals is known. Under electron bombardment and temperatures reached by parts near the cathode and heater, such gases could be evolved steadily, migrating to the electron bombarded areas where they would be partially dissociated, ionized, and accepted into the tubes' ion operating mechanisms.

In Sec. III, the use of the "gas ratio" as an indicator of tube vacuum was mentioned. In making this measurement the control grid is held at cathode potential, electrons are accelerated to the next electrode (positive) and positive ions are collected by the next electrode in the system (held negative). This ion forming mechanism is identical to that used in our positive ion measurements and as they should, the currents check with the total Faraday cage collections closely.

It is obvious that for such gas ratio to represent the space gas, the electron beam must cause space ionization to a greater degree than it causes surface ionization at the electron collecting surface. Such conditions are probably more nearly true for standard ionization gauges where the surface bombardment helps to set a lower limit to gauge sensitivity. However, since our work indicates that the ions formed at the bombarded surface are representative of tube "operation gases," the gas ratio may have value even though it doesn't give residual space gas. Intelligent use of such ratios requires some knowledge of the past history of the tubes being compared.

The cooperation of various members of the General Electric Company's cathode-ray tube division is gratefully acknowledged as is that of Mr. George Parkerson who successfully photographed the ion-burned phosphors.

The Effective Length of an Electrostatic Analyzer*

JAMES E. MONAHAN
Argonne National Laboratory, Chicago, Illinois and St. Louis University, St. Louis, Missouri
(Received October 2, 1952)

The general formula for the effective length of an electrostatic analyzer with "rounded" corners is derived using a modified Schwartz transformation. For the dimensions considered the rounded corner effect introduces a correction of about 0.1 percent in the evaluation of the effective length of the analyzer.

INTRODUCTION

A 90° electrostatic analyzer has been used for some time in conjunction with an electrostatic generator to determine the energy of the ion beam. For an arrangement of diaphragm and deflector plates such as that found in the analyzer, Herzog[1] has shown that the fringing field can be replaced by an equivalent length of an ideal condenser. It is convenient to think of the ideal condenser as being bounded by effective planes outside of which the electric field may be treated as having the value zero, and inside of which, as having the value determined by the geometry of the condenser plates. For an exact evaluation of the energy of the beam it is necessary to determine accurately the position of these planes.

In his calculation of the effective length of an electrostatic condenser, Herzog does not consider the effect of rounding the corners of the deflector and aperture plates. Herb *et al.*[2] found this effect to be negligible (approximately 0.001 percent) for zero aperture slit width and small curvature.

Although the analyzer being designed at Argonne is similar to the one constructed in 1946 at the University of Wisconsin,[2,3] the proposed geometry of the former instrument is such that the zero slit width approximation is no longer applicable. In fact, in order to transmit the greatest possible beam current through the analyzer, the gap in the end diaphragms must be of the same order as the distance between the deflector plates. When well focused, the size of the beam in the Argonne electrostatic generator is about $\frac{1}{8}$ in. to $\frac{1}{4}$ in. in diameter a few feet from the focal position, the focal spot being about $\frac{1}{32}$ in. to $\frac{1}{16}$ in. in diameter. For this reason, a $\frac{3}{16}$-in. gap ($2k$ of Fig. 1) between the deflector plates is chosen as a reasonable minimum value to transmit most of the ion beam; a gap ($2b$) of about $\frac{1}{8}$ in. in the end diaphragms is thus about the maximum usable value without the beam intercepting the deflector plates. When a narrower beam is desired, an adjustable beam defining aperture can be placed beyond these fixed diaphragms without affecting the present calculations. Finally the corners of the deflector plates and end diaphragms must be rounded to radii of the same order as the gap dimensions so that there will not be excessive potential gradients at these positions causing breakdown discharges.

It should be noted also that the end diaphragms used by Herb were assumed to have narrow slit openings and to be infinitely thin. Here they are to be about $8b$ thick with slit openings $2b$. Thus, since the electric field vanishes at a distance of about $4b$ within the diaphragms, it is possible to assume infinitely thick diaphragm plates.

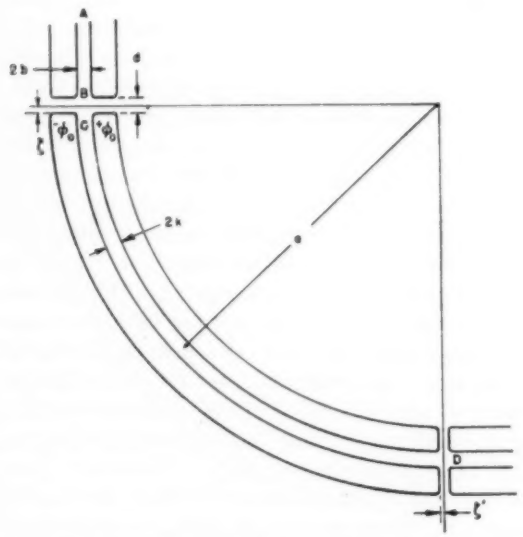


FIG. 1. Outline of the 90° electrostatic analyzer: slits A, B, C, and D are the entrance slit, aperture limiting slit, analyzer slit, and exit slit, respectively. The radius a is the geometric mean of the radii of the analyzer plates, and ζ and ζ' are the distances of the effective entrance and exit planes from the ends of the analyzer.

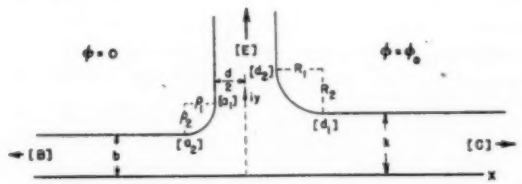


FIG. 2. Plane geometry approximation of the electrostatic analyzer making use of the zero-potential line $y=0$.

* Part of a dissertation submitted for the degree of Doctor of Philosophy at St. Louis University.

[1] R. G. Herzog, Z. Physik **97**, 596 (1935); Arch. Elektrotech. **29**, 790 (1935).

[2] Herb, Snowdon, and Sala, Phys. Rev. **75**, 246 (1949).

[3] Warren, Powell, and Herb, Rev. Sci. Instr. **18**, 559 (1947).

THE SCHWARTZ TRANSFORMATION

As pointed out by Herb, the effect of the curvature of the analyzer plates changes the end correction by a factor $2k/a$ which is ~ 0.01 and hence can be neglected. Furthermore, the aperture plates are grounded and are symmetric with respect to the analyzer so that the geometric symmetry axis is a line of zero potential. Thus, assuming plane parallel plates and making use of the zero potential line, the analyzer can be treated as shown in Fig. 2.

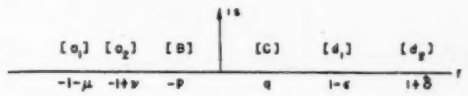


FIG. 3. The "polygon" of the z-plane mapped into the real axis of the t-plane; the related points on the two planes are indicated.

The modified Schwartz transformation[4] for mapping the "polygon" (defined in the z-plane by the partial contour of the analyzer system and the x-axis) on the real axis of the t-plane is

$$z-z_0=A\int\frac{[(t+1+\mu)^{\frac{1}{2}}+\gamma(t+1-\nu)^{\frac{1}{2}}][(t-1+\epsilon)^{\frac{1}{2}}+\lambda(t-1-\delta)^{\frac{1}{2}}]}{(t+p)(t-q)}dt, \tag{1}$$

where λ and γ are real numbers, and A is a constant which may be complex. The transformation parameters, defined in Fig. 3, must satisfy the inequality

$$-1-\mu<-1+\nu<-p<q<1-\epsilon<1+\delta. \tag{2}$$

The constant A is defined by carrying out the integration of Eq. (1) at the point E; for

$$t\gg\begin{cases}1+\delta\\1+\mu,\end{cases}$$

$$d=A\int_{\frown}f(t)dt=i\pi A(1+\lambda+\gamma+\lambda\gamma),$$

where $f(t)$ denotes the integrand of Eq. (1). Thus

$$A=id/(1+\lambda+\gamma+\lambda\gamma)\pi. \tag{3}$$

Integration at the point C as $|t-q|\to 0$ and at the point B as $|t+p|\to 0$ gives

$$\frac{k}{d}=\frac{\pi}{p+q}\sum_{j=1}^{4}\alpha_j(Q_jQ_{j+1})^{\frac{1}{2}}, \tag{4}$$

and

$$\frac{b}{d}=\frac{\pi}{p+q}\sum_{j=1}^{4}\alpha_j(P_jP_{j+1})^{\frac{1}{2}}, \tag{5}$$

where

$$\begin{aligned}
&Q_1=Q_5=1-\epsilon-q, && P_1=P_5=1-\epsilon+p,\\
&Q_2=1+\mu+q, && P_2=1+\mu-p,\\
&Q_3=1+\delta-q, && P_3=1+\delta+p,\\
&Q_4=1-\nu+q, && P_4=1-\nu-p,\\
&\alpha_1=\frac{1}{(1+\lambda+\gamma+\lambda\gamma)\pi}, && \alpha_3=\frac{\lambda\gamma}{(1+\lambda+\gamma+\lambda\gamma)\pi},\\
&\alpha_2=\frac{\lambda}{(1+\lambda+\gamma+\lambda\gamma)\pi}, && \alpha_4=\frac{\gamma}{(1+\lambda+\gamma+\lambda\gamma)\pi}.
\end{aligned} \tag{6}$$

In addition, it is convenient to introduce the following parameters

$$\begin{aligned}
&Q_{12}=2q+\epsilon+\mu, && P_{12}=2p-\epsilon-\mu,\\
&Q_{23}=2q+\mu-\delta, && P_{23}=2p-\mu+\delta,\\
&Q_{34}=2q-\nu-\delta, && P_{34}=2p+\nu+\delta,\\
&Q_{45}=2q+\epsilon-\nu, && P_{45}=2p-\epsilon+\nu,
\end{aligned} \tag{7}$$

in terms of which the integrated transformation equation becomes

$$\begin{aligned}
\frac{z-z_0}{d}=\sum_{j=1}^{4}\alpha_j\Bigg\{&\sin^{-1}\left[\frac{2(t-q)+Q_{jj+1}}{2Q_j-(-1)^jQ_{jj+1}}\right]\\
&+\frac{(Q_jQ_{j+1})^{\frac{1}{2}}}{p+q}\cosh^{-1}\left[\frac{Q_{jj+1}(t-q)-2Q_jQ_{j+1}}{(2Q_j-(-1)^jQ_{jj+1})(t-q)}\right]\\
&-\frac{(P_jP_{j+1})^{\frac{1}{2}}}{p+q}\cosh^{-1}\left[-\frac{P_{jj+1}(t+p)+2P_jP_{j+1}}{(2P_j+(-1)^jP_{jj+1})(t+p)}\right]\Bigg\}.
\end{aligned} \tag{8}$$

The condition that the points $[a_1]$, $[a_2]$, $[d_1]$, and $[d_2]$ in the z-plane map into the corresponding points on the t-plane as indicated in Figs. 2 and 3 gives

$$z_0=ib, \tag{9}$$

$$\begin{aligned}
\frac{\rho_2}{d}=\sum_{j=3,4}\alpha_j\Bigg\{&\cosh^{-1}\left[\frac{2Q_2-Q_{jj+1}}{2Q_j-(-1)^jQ_{jj+1}}\right]\\
&+\frac{(Q_jQ_{j+1})^{\frac{1}{2}}}{p+q}\cos^{-1}\left[\frac{Q_{jj+1}Q_2+2Q_jQ_{j+1}}{(2Q_j-(-1)^jQ_{jj+1})Q_2}\right]\\
&-\frac{(P_jP_{j+1})^{\frac{1}{2}}}{p+q}\cos^{-1}\left[\frac{2P_jP_{j+1}-P_{jj+1}P_2}{(2P_j+(-1)^jP_{jj+1})P_2}\right]\Bigg\},
\end{aligned} \tag{10a}$$

$$\begin{aligned}
\frac{\rho_1}{d}=\sum_{j=1,2}\alpha_j\Bigg\{&-\cos^{-1}\left[\frac{2Q_4-Q_{jj+1}}{2Q_j-(-1)^jQ_{jj+1}}\right]\\
&+\frac{(P_jP_{j+1})^{\frac{1}{2}}}{p+q}\cosh^{-1}\left[\frac{2P_jP_{j+1}-P_{jj+1}P_4}{(2P_j+(-1)^jP_{jj+1})P_4}\right]\\
&-\frac{(Q_jQ_{j+1})^{\frac{1}{2}}}{p+q}\cosh^{-1}\left[\frac{2Q_jQ_{j+1}+Q_{jj+1}Q_4}{(2Q_j-(-1)^jQ_{jj+1})Q_4}\right]\Bigg\},
\end{aligned} \tag{10b}$$

[4] W. R. Smythe, *Static and Dynamic Electricity* (McGraw-Hill Book Company, Inc., New York, 1939), p. 97.

$$\frac{R_1}{d}=\sum_{j=2,3}\alpha_j\left\{-\cos^{-1}\left[\frac{2Q_1+Q_{jj+1}}{2Q_j-(-1)^jQ_{jj+1}}\right]\right.$$
$$+\frac{(Q_jQ_{j+1})^{\frac{1}{2}}}{p+q}\cosh^{-1}\left[\frac{2Q_jQ_{j+1}-Q_{jj+1}Q_1}{(2Q_j-(-1)^jQ_{jj+1})Q_1}\right]$$
$$\left.-\frac{(P_jP_{j+1})^{\frac{1}{2}}}{p+q}\cosh^{-1}\left[\frac{2P_jP_{j+1}+P_{jj+1}P_1}{(2P_j+(-1)^jP_{jj+1})P_1}\right]\right\}, \quad (11a)$$

$$\frac{R_2}{d}=\sum_{j=1,4}\alpha_j\left\{\cosh^{-1}\left[\frac{2Q_3+Q_{jj+1}}{2Q_j-(-1)^jQ_{jj+1}}\right]\right.$$
$$+\frac{(P_jP_{j+1})^{\frac{1}{2}}}{p+q}\cos^{-1}\left[\frac{2P_jP_{j+1}+P_{jj+1}P_3}{(2P_j+(-1)^jP_{jj+1})P_3}\right]$$
$$\left.-\frac{(Q_jQ_{j+1})^{\frac{1}{2}}}{p+q}\cos^{-1}\left[\frac{2Q_jQ_{j+1}-Q_{jj+1}Q_3}{(2Q_j-(-1)^jQ_{jj+1})Q_3}\right]\right\}. \quad (11b)$$

The form of the function is determined in each case by the range of its argument from the inequalities (2) using the following relations:

$t<-1$;

$$\cos^{-1}t=\pi-i\cosh^{-1}|t|,$$
$$\sin^{-1}t=-\frac{\pi}{2}+i\cosh^{-1}|t|,$$
$$\cosh^{-1}t=i\pi+\cosh^{-1}|t|,$$

$-1<t<0$;

$$\cos^{-1}t=\pi-\cos^{-1}|t|,$$
$$\sin^{-1}t=-\sin^{-1}|t|,$$
$$\cosh^{-1}t=i\pi-i\cos^{-1}|t|,$$

$0<t<1$;

$$\cosh^{-1}t=i\cos^{-1}t,$$

$t>1$;

$$\cos^{-1}t=-i\cosh^{-1}t,$$
$$\sin^{-1}t=\frac{\pi}{2}+i\cosh^{-1}t.$$

These follow immediately from the condition that the transformation function be chosen such that as t increases, the polygon in the z-plane is traversed in a counterclockwise sense.

If the modified Schwartz transformation is to have any practical significance, it is necessary that Eq. (8) map into a physically reproducible contour in the region of the "rounded corners." The radii of curvature are: for $1-\epsilon<r<1+\delta$,

$$\frac{R(r)}{d}=\left[\left(\frac{1}{2}+\frac{R_1}{d}-\frac{x(r)}{d}\right)^2+\left(\frac{k}{d}+\frac{R_2}{d}-\frac{y(r)}{d}\right)^2\right]^{\frac{1}{2}}, \quad (12)$$

and for $-1-\mu<r<-1+\nu$,

$$\frac{\rho(r)}{d}=\left[\left(\frac{1}{2}+\frac{\rho_1}{d}+\frac{x(r)}{d}\right)^2+\left(\frac{b}{d}+\frac{\rho_2}{d}-\frac{y(r)}{d}\right)^2\right]^{\frac{1}{2}}. \quad (13)$$

In all cases considered (about twenty points) it has been found that by setting $R_1=R_2$ and $\rho_1=\rho_2$ and solving the resulting equations for λ and γ, R and ρ are make independent of the auxiliary variable r within the ranges indicated; i.e., the analyzer "corners" are the arcs of circles within the accuracy of the calculations, i.e., about 2 percent.

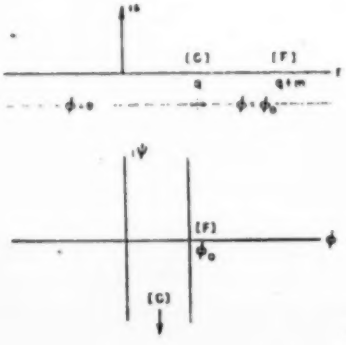


FIG. 4. The known field of an infinite condenser on the x-plane mapped on the upper half of the t-plane.

THE EFFECTIVE LENGTH OF THE ANALYZER

To determine the effective length of the analyzer system it is necessary to evaluate the field at every point along the path of the ion beam. This can be done by using Eq. (8) to eliminate the auxiliary variable t from the transformation which maps the known field of an infinitely extended condenser on the upper half of the t-plane. The Schwartz theorem gives for the latter transformation (Fig. 4):

$$\chi=\varphi+i\psi=\varphi_0\left[1+\frac{i}{\pi}\ln(t-q)\right]$$
$$=\frac{\varphi_0}{\pi}\left\{\tan^{-1}\frac{s}{q-r}+i\ln[(q-r)^2+s^2]^{\frac{1}{2}}\right\}, \quad (14)$$

where φ is the potential function, ψ is the stream function, and $2\varphi_0$ is the potential difference of the analyzer plates.

Since the first-order focusing theory developed for the 90° electrostatic analyzer is concerned with the motion of particles in the neighborhood of the x-axis, it is possible to expand Eq. (14) neglecting quadratic and higher order terms in s. Thus for $-p<r<q$,

$$\psi=\frac{\varphi_0}{\pi}\ln(q-r). \quad (15)$$

For a particle of mass m and charge e which enters a field E_y along the x-axis with a uniform velocity v_x, the angle of deflection α at any point in the field is given by

$$\tan\alpha=\frac{e}{mv_x^2}\int_{-\infty}^{x}E_y dx=\frac{e}{mv_x^2}[\psi(x)-\psi(-\infty)]. \quad (16)$$

where the effect of the small tangential component of the end-field has been neglected. Thus in the neighbor-

hood of the symmetry axis

$$\tan\alpha=\frac{e}{mv_x^2}\frac{\varphi_0}{\pi}\ln\frac{q-\zeta}{q+p}, \tag{17}$$

since $r=p$ is the image of the point $x=-\infty$. The analogous expression for an ideal condenser where the constant field begins at the effective plane

$$x=\frac{d}{2}-\zeta$$

is

$$\tan\bar{\alpha}=-\frac{e}{mv_x^2}\int_{d/2-\zeta}^{x}\frac{\varphi_0}{k}dx=\frac{-e\varphi_0}{mv_x^2k}\left(x+\zeta-\frac{d}{2}\right), \tag{18}$$

where ζ is so defined that for large values of x, α and $\bar{\alpha}$ are equal.

To obtain α as a function of x, the real part of Eq. (8) is solved for $\ln(q-r)$ and substituted into Eq. (17). The resulting expression is then equated to the right-hand side of Eq. (18) to obtain

$$\frac{\zeta}{d}=\frac{1}{2}-\sum_{j=1}^{4}\alpha_j\left\{\sin^{-1}\left[\frac{Q_{jj+1}}{2Q_j-(-1)^jQ_{jj+1}}\right]+\frac{(Q_jQ_{j+1})^{\frac{1}{2}}}{p+q}\ln\frac{4Q_jQ_{j+1}}{(2Q_j-(-1)^jQ_{jj+1})(p+q)}-\frac{(P_jP_{j+1})^{\frac{1}{2}}}{p+q}\cosh^{-1}\left[\frac{P_{jj+1}(p+q)+2P_jP_{j+1}}{(2P_j+(-1)^jP_{jj+1})(p+q)}\right]\right\}. \tag{19}$$

In the derivation of Eq. (19) use has been made of the approximation $\cosh^{-1}z\sim\ln 2z$ for $z\gg 1$.

Equations (4), (5), (10), (11), and (19) relate the equivalent length to the physical dimensions of an analyzer system by means of the eight transformation parameters defined in Fig. 3. Although it is practically impossible to eliminate these parameters to obtain ζ in terms of the analyzer dimensions, two circumstances combine to make possible the use of the parametric form of these relations.

1. The above set of equations involves only ratios of analyzer distances, several of which can be adjusted physically within certain limits to fit the results of a calculation.

2. By a careful examination of the transformation itself, as well as the resulting equations, it is possible to choose values for the parameters so as to obtain the end correction for analyzer ratios within the limits mentioned in item 1.

NUMERICAL METHODS

The equivalent length of an ideal analyzer for the case in which only the edges of the deflector plates are rounded is given by the foregoing equations with $\mu=\nu=\gamma=0$. Of particular interest, however, is the system, treated by Herzog,[1] in which all corners are square. In this case it is possible to solve for the transformation parameters p and q in terms of the analyzer distances. Setting $\mu=\nu=\gamma=\epsilon=\delta=\lambda=0$, and solving Eqs. (4) and (5) for p and q, one obtains

$$p'=\frac{d^2+k^2-b^2}{[(d^2+k^2+b^2)^2-4k^2b^2]^{\frac{1}{2}}}, \tag{20}$$

$$q'=\frac{d^2-k^2+b^2}{[(d^2+k^2+b^2)^2-4k^2b^2]^{\frac{1}{2}}}, \tag{21}$$

where the primes distinguish between parameters referring to square and rounded-corner transformations. Since the rounding of the corners introduces only a first-order correction, it is possible to use Eqs. (20) and (21) to obtain a zero-order approximation for p and q which will give desired ratios of b/d and k/d.

The problem of choosing parameters to obtain given values of analyzer dimensions is considerably simplified by the following considerations. The range of the transformation parameters is limited by Eq. (2). Further, Eqs. (10) and (11), together with the above described method for evaluating λ and γ, result in the limits

$$\rho(\nu\to 1-p)=\infty, \quad \rho(\mu\to\infty)=\infty, \tag{22}$$

$$R(\epsilon\to 1-q)=\infty, \quad R(\delta\to\infty)=\infty, \tag{22a}$$

where ρ is a slowly varying function of δ, ϵ and R a slowly varying function of μ, ν. An increase in the value of μ or δ tends to increase the ratios k/d and b/d.

For $R>\rho$ and given values of k/d and b/d, the value of p' predicted by Eq. (20) is less than the value of p necessary to obtain the same values of these analyzer ratios; on the other hand Eq. (21) yields a value of q' which is greater than the corresponding value of q. For $\rho>R$, the reverse is true. Also from Eqs. (4) and (5)

$$b(p\to 1)=0, \quad k(q\to 1)=0. \tag{23}$$

The actual steps in constructing the analyzer are to make the plates to the chosen mean curve of radius a and to find empirically the value $2k$ which is most nearly constant. It is then possible, for a given set of transformation parameters, to determine d and all other dimensions, b, R, ρ, and ζ, in terms of the measured value of k. It is hoped that the dimension $2k$ will be constant to 0.0002 in. out of approximately 0.19 in.,

TABLE I. Transformation parameters and corresponding analyzer dimensions.

	p	q	ϵ	δ	ν	μ	λ	γ	k/d	b/d	R/d	ρ/d	ζ/d
1	0.75	0.25	0.70	10.00	0.20	3.00	0.26	2.66	1.12	1.01	0.92	0.62	0.25
2	0.85	0.25	0.70	10.00	0.10	1.50	0.27	2.28	0.97	0.78	0.85	0.44	0.15
3	0.94	0.25	0.71	12.00	0.01	1.50	0.24	2.15	0.92	0.69	0.90	0.42	0.10
4	0.94	0.25	0.71	14.00	0.01	1.50	0.23	2.20	0.95	0.76	0.94	0.42	···
5	0.94	0.20	0.76	14.00	0.01	1.50	0.23	2.11	0.98	0.78	0.96	0.42	0.09
6	0.95	0.10	0.88	20.00	0.01	1.00	0.18	2.10	0.93	0.70	1.14	0.38	0.07
7	0.95	0.05	0.93	20.00	0.01	1.00	0.18	1.89	0.94	0.71	1.16	0.42	0.04
8	0.96	0.00	0.98	20.00	0.01	1.00	0.18	2.19	0.96	0.72	1.16	0.42	0.03

and that the effective mean $2k$ will be known to better than 0.0001 in. This is an accuracy of about 0.05 percent. Hence if all other corrections can be held to about 0.01 percent, it may be possible to determine the absolute voltage scale at high voltages to about twice the accuracy already achieved (0.1 percent).

An accuracy of 0.01 percent in analyzer length for a 40-in. radius is about 0.006 in. For the proposed value of $2k$, the value of d is approximately 0.1 in. for all ratios given in Table I. Thus a length correction, ζ/d, less than 0.06 can be neglected. In order to obtain maximum ion current, it is necessary that b/k be greater than or equal to $2^{-\frac{1}{2}}$; and finally, as mentioned before, the radii R and ρ must be of the order of the gap dimensions. With the exception of p and q, which can be evaluated approximately from Eqs. (20) and (21), the parameters listed in row 1 (Table I) are assigned arbitrarily. However, once this initial calculation is made, the behavior of the analyzer dimensions with respect to the transformation parameters described by Eqs. (22) and (23) may be used to obtain values of the analyzer ratios successively closer to any desired set. The ratios given in row 8 are within the limits of the proposed dimensions of the present analyzer.

It is interesting to note that for the system described by the last line of the table ($\zeta/d=0.03$), the end correction neglecting the rounded corners gives $\zeta/d=0.37$. This would result in an error somewhat greater than 0.05 percent in the determination of the equivalent length of the analyzer.

JOURNAL OF APPLIED PHYSICS VOLUME 24, NUMBER 4 APRIL, 1953

Electrical Phenomena in Adhesion. I. Electron Atmospheres in Dielectrics*

SELBY M. SKINNER, ROBERT L. SAVAGE, AND JOHN E. RUTZLER, JR.
Department of Chemistry and Chemical Engineering, Case Institute of Technology, Cleveland 6, Ohio
(Received December 11, 1952)

Rapid breaks of metal-polymer-metal adhesive specimens have shown the presence of a charge density on the metallic surface, provided the break occurs at the metal-polymer interface, and a much smaller charge if it occurs in the interior of the polymer. This is analyzed in terms of the electron atmosphere existing external to the metal in a dielectric region of low barrier. The barrier values in certain cases are otherwise known to be of appropriate order of magnitude. Measured experimental surface charge densities permit estimation of the maximum limit of barrier height. Because of the presence of the charge distribution in the polymer, there is an electrostatic force on the adherend metal directed toward the dielectric polymer, which represents a contribution to the total adhesive forces opposing break. Such electrostatic contributions to adhesion have not previously been considered; their order of magnitude and their relation to the thickness of the adhesive are discussed. The qualitative agreement between the theory and a number of previously unexplained experimental results in the literature is shown.

I. INTRODUCTION

IN the course of a fundamental investigation of adhesion and the phenomena contributing to it, it has been observed that when adhesive specimens, Fig. 1, consisting of metal plates bonded by high polymer adhesives are broken in tensile test, electrical charges are found on the two separating halves. If the opposite charges on the two halves are permitted to flow through an external circuit and the IR drop is recorded by a cathode-ray oscillograph, traces such as those in Fig. 2 are obtained. The experimental arrangements and results will be described in detail in the second article of this series. It is the purpose of the present paper to consider the possible origin of the charges, how the charge distribution may be related to experimental parameters, and what relation the charges may have to adhesion. Accordingly, for the present purpose, the only experimental facts resulting from this investigation which are utilized are:

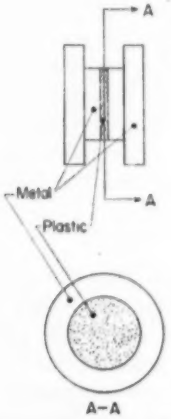


FIG. 1. Adhesive specimen.

1. There is a charge distribution in the polymer if the actual break occurs in a reasonably short time; the period of stressing before break may vary between wide limits, however, without alteration of the essential behavior.

2. The charge density is considerably greater if the polymer is stripped from the metal (adhesive break),

* With the support of the Flight Research Laboratory, Wright Air Development Center, U. S. Air Force, Contract AF33(038)-26461.

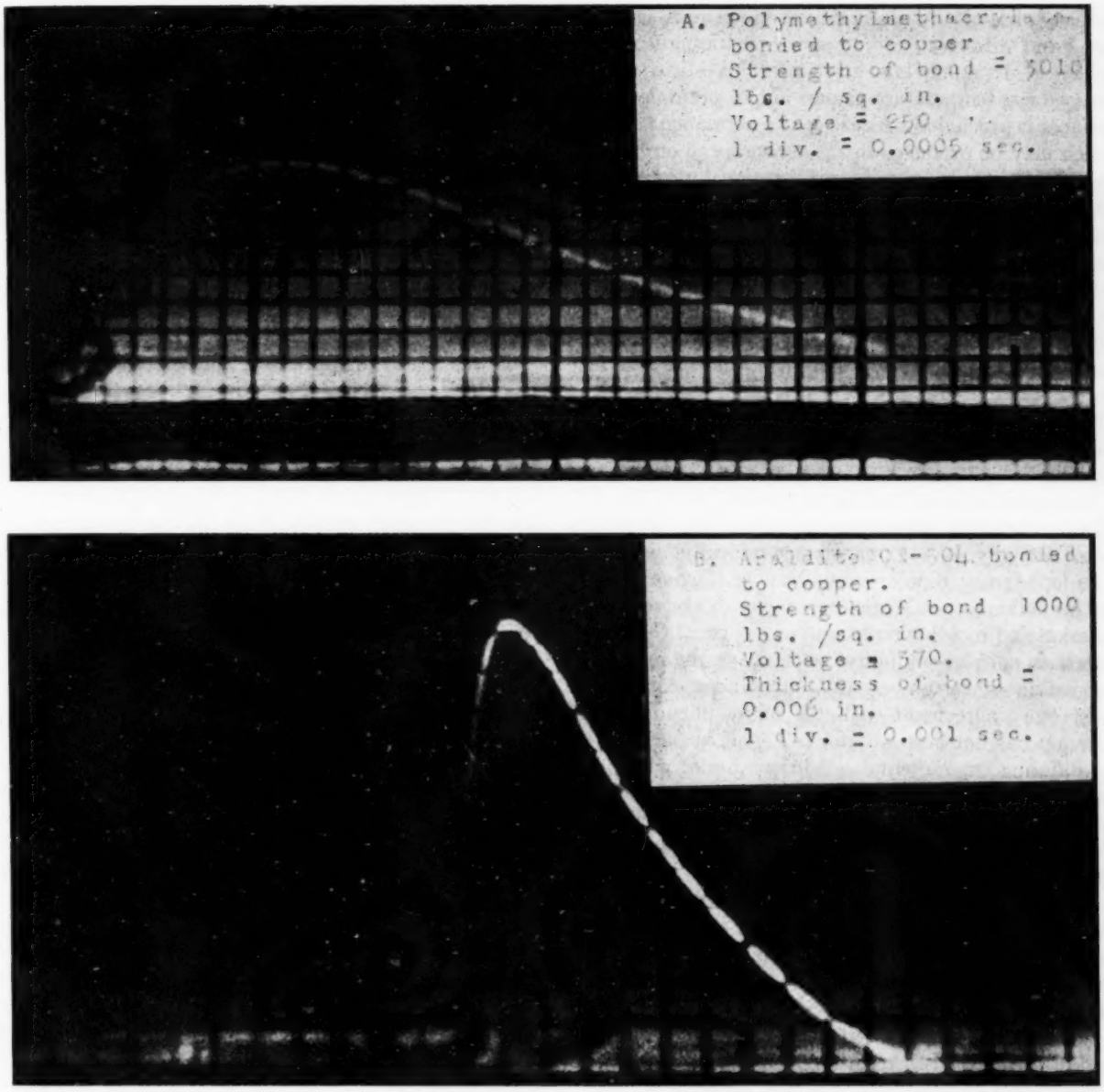

A. Polymethylmethacrylate bonded to copper
Strength of bond = 5010 lbs. / sq. in.
Voltage = 250
1 div. = 0.0005 sec.
B. Araldite 103-504 bonded to copper.
Strength of bond 1000 lbs. /sq. in.
Voltage = 570.
Thickness of bond = 0.006 in.
1 div. = 0.001 sec.

FIG. 2. Oscillograph traces of potentials.

than if the break occurs in the body of the polymer (cohesive break). Even in the latter case, charge densities of low magnitude are observed.

3. The sign of the charge on each of the two halves of the specimen depends upon the metal and the polymer used in preparing the adhesive bond. The magnitude of the charge varies with experimental conditions, and breaks have been obtained in which the observed charge density has apparently been limited only by the breakdown potential of the air in the process of separating the plates of the specimen.

4. The phenomena in 1–3 above have been observed with a number of different adhesives, and with various metals bonded to them.

Adhesion is a complex phenomenon, depending upon the simultaneous action of a number of different mechanisms. In industrial practice, empirical methods are utilized to obtain the right amount of adhesion combined with other desired properties of the finished materials. In the laboratory, it is found that the number of parameters and possible mechanisms which must be considered requires the application of painstaking surface chemistry techniques, and care in minute procedural detail, if the results are to be related to a single cause rather than a mixture of various causes. The fact that adhesion involves an interface between two different substances introduces a degree of complexity not found in investigations of cohesive strengths. Not only do the two surfaces to be joined possess the forces which distinguish a surface from the interior of the

material, but also local differences in chemical composition, and adsorbed and other impurities affect their behavior. If nonmetals are under investigation, the use of high temperatures as an aid to obtaining clean surfaces is precluded because of the permanent changes which may be produced in the substance upon heating. In such a case, a specific preparational procedure is adopted, in order that successive specimens may have reproducible (though not necessarily specifiable) surface characteristics.

On the other hand, strong, reproducible bonds can be obtained quite regularly with suitable techniques. This would not be possible, if the practice and research in adhesives were entirely empirical, and the interface were a random mixture of unknown impurities. The adhesive bond is apparently less likely to suffer from such random factors than is the free surface before bonding.[1]

In general, the mechanisms employed to explain the adhesion between substances have been van der Waals forces (broadly defined), or, more specifically, dispersion (London) forces, dipole or quadrupole (Debye) forces, and induction forces. While these forces can be related to cohesion and to adhesion as well as to other behavior of materials, such as solubility, swelling, or the molecular association of certain organic substances, completely satisfactory agreement with the many phenomena of adhesion has not been obtained. This is in part due to the differing experimental conditions under which adhesion data have been obtained, and in part to the complex set of mutually operating phenomena involved in everyday adhesion. Experimental methods are required which identify and disentangle such factors as the role of the solvent and solvent concentration in the adhesive, the effect of the bulk elastic properties of the adhesive, the effect of surface structure of the adherend, or the extent to which the dipole concept, for example, may substitute for a direct consideration of the charge density distributions and potential energy in the neighborhood of the interface. Accordingly it becomes desirable to investigate the phenomena which occur at adhesive interfaces, and systematically to examine them for their applicability and relation to the phenomenon of adhesion. The electric charge accompanying the breaking of the sample is one such phenomenon.

II. THE EXPERIMENTAL BEHAVIOR

Visualizing the adhesive specimen (Fig. 1) before and after break, and without regard to the physical nature of the adhesion before break, it is evident that any charge which may remain on the surfaces after break is directly related to a charge distribution analogous to that in a parallel plate condenser.[2] It is possible to determine the total charge and often the corresponding area of the plates, and thus to determine the charge density. However, adhesives do not cooperate with the experimenter, and often the break is in the body of the adhesive, or both plates show clean metal portions and portions covered with adhesive. An independent verification of the charge density is desirable, and this can be done by suitable experimental techniques.

The metal surfaces were bonded with high polymer adhesives either by prepolymerization and application, or by polymerization in place. After a number of preliminary measurements, the sample was broken by a modification of the ASTM test method[3]; electrical contact was made to the plates and the charge resulting from the break passed through a high resistance shunted by a cathode-ray oscilloscope. Since the plates are of small mass, the large forces applied to produce break cause rapid separation after break. This magnifies the potential difference E_0 between them; at the same time, the flow through the external, measuring, circuit decreases E_0. When the separation is sufficiently rapid, an oscillograph trace with a maximum is obtained, the maximum being from ten to a thousand times E_0. As shown in another paper, analysis of the trace permits determination of the charge density initially left on the metal surface by separation of the adhesive.[4]

Since under appropriate experimental conditions, the whole oscilloscope trace is obtained in from one to fifteen milliseconds, and the portion of the trace concerned with the removal of the whole area of the adhesive[5] is of the order of 10^{-4} second or considerably less, the experimental data are obtained before surface contamination or other effects can produce markedly changed conditions of the surface.

The existence of the charge at break can be explained qualitatively in several ways, according to the physical limitations imposed. Accepted knowledge of the solid state supplies what appears to be the most reasonable explanation. The sample consists of a dielectric bonded to metal plates on either side. The metal plates have a characteristic Fermi distribution of electrons, the dielectric has a characteristic set of energy states modified by impurities, surface states, and proximity to the metal; between the two there is a barrier which depends upon the specific characteristic quantities of each. Electrons or holes may be transferred between metal and dielectric. In the dielectric there will occur a distribution of charge dependent upon the barrier between them; if this is a negative charge, it is equivalent to an electron atmosphere. The break is made in a time of the order of, or smaller than, the relaxation time of the

[1] H. A. Perry, Jr., presented at the Symposium on Adhesion, Case Institute of Technology, April 24–5, 1952.

[2] The analogy is not perfect. This will be treated in a succeeding paper.

[3] ASTM Method of Test for Tensile Properties of Adhesives, No. D-897-49.

[4] Surface charge density, and not the initial electrostatic potential between buttons is the object of the analysis.

[5] The removal of the adhesive from the metal involves a succession of breaking of individual bonds between successive surface groups in the adhesive and metal. Obviously the time required for the rupture of an individual bond is a small fraction of the total time taken to strip the adhesive film from the metal.

electrons in the atmosphere which has been established in the dielectric external to the metal. Accordingly, charges remain in the dielectric and opposite charges are measured on the metal which has been pulled away from it. The sign of the charge depends upon the particular mechanism responsible, and upon whether the barrier is positive or negative. The break mechanically separates the metal from the dielectric against all forces acting, with, in favorable cases, negligible presumed disturbance of charge density.[6] Thus, the surface charge observed is directly related to the volume density of charge in the dielectric, which in turn is directly related to what amounts to an effective work function between the two substances. The relationships will be considered below assuming the transfer of electrons,[7] and the experimental results will be covered subsequently.

The existence of the charge in the dielectric of the original sample before break, and the corresponding opposite charge on the metal show that part of the adhesive force in metal-polymer adhesive pairs is electrostatic. Such a force has not previously been included in the analysis of adhesion. The possible significance is discussed here.

III. THE PHYSICAL PICTURE

Since charge was transferred, the various van der Waals mechanisms are not suitable mechanisms for the discussion. A possible interpretation is one involving chemical bonding between the metal and the polymer, such as the conversion of a double bond in a monomer unit into a single bond with the other bond attached to the metal. Under such conditions, a more stable state may result if the electron from the metal moves into the monomer structure, with an eventual increased electron density about the carbon atom which originally was double bonded. However, the picture is too simple since it implies that the energy levels in the solid are equivalent to those in the free monomer.

Looked at quantum mechanically, the electrons entering the polymer will enter describable states, whose specification must conform to the fact that the polymer as a whole was electrically neutral before contact with the metal. The essential difference between a metal and an insulator is that the metal has an energy band only partially filled with electrons, so that with small applied fields transitions to conducting states are possible, whereas in the insulator, there is a discrete energy gap separating two bands, the higher of which is empty at low temperatures and the next lower filled. At ordinary temperatures this gap cannot be traversed, but at high temperatures, sufficient thermal energy is available to permit electrons to make such transitions, and, for example, salts become conducting.

Considerable theoretical and experimental investigation has been devoted to establishing the details of the metallic state. Less has been done with nonmetals, except in the case of ionic crystals and semiconductors. Since the adhesives considered here are high polymers, much less is known about their detailed structure from the quantum-mechanical point of view. X-ray results show that some high polymers exhibit a microcrystalline structure below the second-order transition point, and that within the crystallite, there is a certain quasi-periodicity of structure due to the statistical linking of similar monomer units. On the other hand, this structure is far from the regular lattice exhibited by metals or regular crystals. Not only is the molecular weight of the linked monomer units a statistical distribution determined by the course of the polymerization and the various chain-terminating mechanisms, but also there may exist within the mass debris of the nature of branched and cross-linked chains, and, depending upon the method of polymerization, traces of the polymerizing agent or catalyst, or sometimes even small concentrations of electrolyte. Insufficient information is available with respect to a particular sample of adhesive, no matter how carefully prepared, to permit detailed calculations of the distributions of energy levels and the degree to which they are filled, the variation of atomic charge density throughout the adhesive, or the contributions which may be expected from impurities, surface or interface states, etc. For the present purpose, it will be sufficient to treat in detail a picture representing the polymer as a region in which excess electrons of sufficient energy can enter conduction levels[8] and move freely with a uniform potential energy. This approximation corresponds to that made in the Drude-Sommerfeld electron theory of metals. Certain refinements of this concept will then be discussed.

When the barrier for the passage of electrons between a metal and a contiguous material is of the order of a few tenths of an electron volt or less, an electron atmosphere of considerable magnitude may exist in the latter substance at room temperatures. Such phenomena are not normally observed in air surrounding metals, since the potential barrier for evaporation of electrons into air or vacuum is of the order of several electron volts, and only at high temperatures is the thermal energy sufficient to permit an appreciable fraction of the electrons to exist in the vapor external to the metal.

[6] The measured charge density is a lower limit to the actual charge density which existed in the dielectric prior to break, since if local fields on the surfaces exceed the values at which corona discharge into the air occurs, a portion of the charge is lost, and since, if the total area of adhesive break is not so great as estimated, the charge density is actually greater. Care is taken to avoid, so far as possible, such high fields. For the analysis, corona discharge into air during separation will be left out of consideration. This will be taken up in the next paper.

[7] Analogous considerations are of course possible, if the sign of the charge, or the direction of transfer is opposite to that chosen here.

[8] Excited states in the insulator which are nonconducting are not considered, since in such states the electron would remain in the vicinity of a particular atom, whereas the experiments show that there is often a very small but measurable charge on the surfaces resulting from break within the interior of the adhesive.

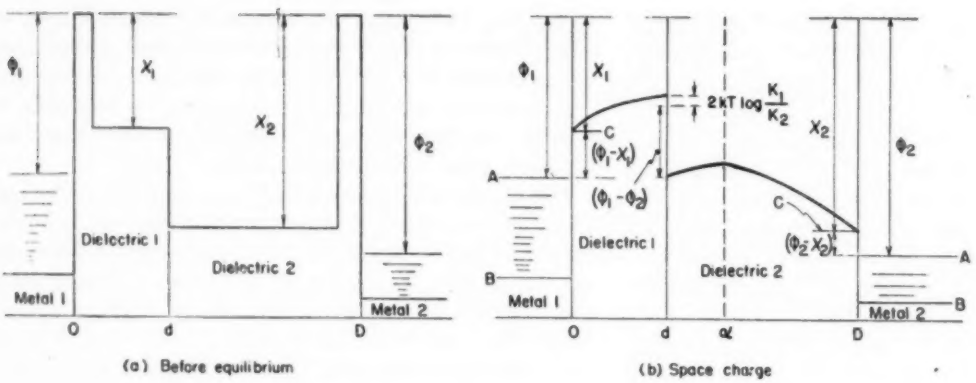


FIG. 3. Barrier potentials on contact between a dielectric and a metal.

If the contiguous surface is another metal, separation of the two permits observation of a contact potential. Similarly, if a nonmetal is brought into contact with a metal, potentials may be measured; complete investigation is more difficult because dielectrics are relatively nonconducting, and surface charges on insulators behave erratically. A momentary contact between an insulator and a metal is very different from an aged adhesive bond; leaving aside chemical interaction, the former results in two interfaces with a thin film of gas between the substances. For the electron, this constitutes an additional barrier, not existing in the intimate contact achieved at an adhesive interface; at the latter, the barrier may be very low or even negative. Certain consequences of a very low barrier are very different from those normally expected in ordinary experience. In the air, or in contact with another dielectric, an adhesive substance, even with considerable "electron affinity," may persist in a relatively neutral state, because of the small availability of electrons. When bonded to a metal, however, it is in contact with a reservoir of electrons, whose diffusion into the adhesive takes place under conditions similar to those involved in thermionic phenomena. The charge which "evaporates" into the adhesive is chiefly retained in the region nearest to the metal, but under certain conditions appreciable charge densities may exist in the body of the adhesive.

IV. THE ELECTRON ATMOSPHERE

Let χ be the work required to remove an electron from the lowest state in the *conduction band* in the insulator to infinity; the potential energy of the free electrons in the insulator is then $-\chi$, and χ may be called the "electron affinity"; similarly let ϕ be the work function of the metal. When the insulator is some distance from the metal, the Fermi level of the electrons in it will be different from that in the metal. If the two substances are brought into contact, electrons flow from the higher levels in the one to the lower, unoccupied levels in the other. The exact description depends upon the nature of the barrier; here, a step function will be sufficient [Fig. 3(a)]. The flow of electrons alters the occupied Fermi levels in the two until a minimum overall energy is obtained when the Fermi levels are at the same height. The insulator can carry current (and therefore accumulate charge in its interior) only if electrons are in the conduction levels. The necessary energy may be obtained in several ways, one of which is temperature agitation. If the barrier is large, insufficient energy is gained in this way except at high temperatures; with a low barrier, an appreciable fraction will have sufficient energy to enter the conduction levels at room temperature. This fraction, from the expression for the Fermi-Dirac distribution, will be zero at 0°K and increase with temperature. As the electrons pass into the dielectric, a space charge is formed, so that the potential seen by subsequent electrons is successively higher. At equilibrium, a barrier similar to that in Fig. 3(b) will exist.[9]

Consider two metal surfaces separated a distance D in the direction of x, with two distinct dielectrics occupying the space between them. It is assumed that all quantities depend only on the x coordinate. The field approaches zero and the potential energy V increases logarithmically, as x becomes large.[10] For a point in

[9] The following references will, in the text, be referred to by the names of the authors. R. H. Fowler, *Statistical Mechanics* (Cambridge University Press, Cambridge, 1936), second edition; R. H. Fowler and E. A. Guggenheim, *Statistical Thermodynamics* (Macmillan and Company, Ltd., London, 1939); N. F. Mott and R. W. Gurney, *Electronic Processes in Ionic Crystals* (Oxford University Press, London, 1940), especially Chapter V; F. Seitz, *Modern Theory of Solids* (McGraw-Hill Book Company, Inc., New York, 1940). Mott and Gurney particularly review the evidence for the existence of low or negative potential barriers at the contact between a metal and an insulator. Seitz gives a fairly detailed account of surface states.

[10] The derivation is given in N. F. Mott and R. W. Gurney (reference 9), pp. 170–1; see also R. H. Fowler (reference 9), p. 367. Essentially it is: Letting Ne be the volume density of charge in the dielectric as a function of x, where e is the charge of the electron, v and D the mobility and diffusion constant of the electrons, the condition that the electron atmosphere in either dielectric has reached equilibrium is that no current is flowing. Therefore $NveF-eDdN/dx=0$. (For distances so small that the electron has small probability of collision, this relation is not strictly true; however, to the order of approximation being used, it may be accepted.) With the Einstein relation $vkT=eD$, this integrates into $\log(N/N_0)=e/kT\int_0^x F dx$, in which N_0 is the density of electrons in the insulator at the contact with the metal surface. Upon passing into the interior of the insulator $N(x)$ will decrease. The positive charge per unit area on the metal will correspond to the electrons lost by that area of the metal into the interior of the dielectric. The potential energy of an electron at a distance x from the metal surface is $V(x)=-e\int_0^x F dx$, where F is the field in the insulator; F obeys the space charge relation $dF/dx=4\pi Ne/K$. Integration of the equations subject to the condition mentioned above yields Eq. (1).

either space, integration of the space charge equations gives

$$V=2kT\log(ax+b), \tag{1}$$

where a and b are constants to be determined by the boundary conditions.

Figure 3(b) pictures the energy levels; the Fermi levels are not in question in the dielectric. $\phi-\chi$ is the energy gap between the highest occupied levels in the metal and the lower edge of the conducting band in the dielectric. Since the work functions of the two metals may be different, and the depth of the conduction levels in the two dielectrics may also differ, the boundary conditions to be satisfied are (K being the dielectric constant),

$$\left.\begin{aligned} &V(0)=-\chi_1;\quad V(D)=-\chi_2\\ &\lim_{\epsilon\to 0}\left[K_1F(d-\epsilon)-K_2F(d+\epsilon)\right]=0\end{aligned}\right\} \tag{2}$$

plus two additional conditions, which become obvious if the simpler case of the usual adhesive bonded specimen is considered. In this special case, there is only one dielectric, the plates are identical, and obviously the field is symmetrical about the center plane, being directed towards the two surfaces. In the present case, there is no pressing reason to expect symmetry, but, for not too great a disparity between the "emissivities" of the metal plates, there certainly is a value of x within the dielectrics for which the field changes direction. Calling this value α, and denoting by g the ratio of the absolute values approached by the (oppositely directed) fields from the left and the right of α, the conditions are

$$\left.\begin{aligned} &\lim_{\epsilon\to 0}\left[V(\alpha+\epsilon)-V(\alpha-\epsilon)\right]=0\\ &\lim_{\epsilon\to 0}\left[F(\alpha+\epsilon)+gF(\alpha-\epsilon)\right]=0\end{aligned}\right\}. \tag{3}$$

Two cases must be considered, namely, those in which $\alpha\leqslant d$, and in which $\alpha\geqslant d$; that is, when the change in field direction occurs in the first dielectric and when it occurs in the second.

Solution is made for V and from it are obtained F and N by the relations

$$F=-\frac{1}{e}\frac{dV}{dx};\quad N=\frac{K}{4\pi e}\frac{dF}{dx}.$$

The value of g must still be fixed by a reasonable assumption. Since the metal surfaces may be regarded as sources of electrons diffusing outwards into the dielectric, the location of α will depend upon the relative effectiveness of the plates in furnishing electrons, i.e., upon the ratio of the charge densities from the two. At the point at which the two charge densities from each considered separately are equal, a displacement into the dielectric in either direction will cause the contribution of the plate which is approached to become dominant, and that from the other plate to become less. The point at which the field changes direction is therefore that at which the separate charge densities from each plate are equal. When expressed mathematically in terms of the expressions obtained for the charge densities, it is found that g can only be ±1. The value -1 is rejected since it is equivalent to the simple case treated by Mott and Gurney. With the value $g=+1$, the solutions are:

For $\alpha\leqslant d$

Region	V	$N(x)$
$0\leqslant x\leqslant\alpha$	$-\chi_1+2kT\log(1+\xi_1)$	$N_{01}(1+\xi_1)^{-1}$
$\alpha\leqslant x\leqslant d$	$-\chi_1+2kT\log\eta$	$N_{01}\eta^{-1}$
$d\leqslant x\leqslant D$	$-\chi_2+2kT\log(1+\xi_2)$	$N_{02}(1+\xi_2)^{-1}$

For $\alpha\geqslant d$

Region	V	$N(x)$
$0\leqslant x\leqslant d$	$-\chi_1+2kT\log(1+\xi_1)$	$N_{01}(1+\xi_1)^{-1}$
$d\leqslant x\leqslant\alpha$	$-\chi_2+2kT\log\zeta$	$N_{02}\zeta^{-1}$
$\alpha\leqslant x\leqslant D$	$-\chi_2+2kT\log(1+\xi_2)$	$N_{02}(1+\xi_2)^{-1}$

(4)

where

$$\left.\begin{aligned} &\xi_1=\frac{x}{x_{01}};\quad \eta=\frac{K_2(d-x)+K_1(D-d+x_{02})}{K_2x_{01}};\\ &\xi_2=\frac{(D-x)}{x_{02}};\quad \zeta=\frac{K_1(x-d)+K_2(d+x_{01})}{K_1x_{02}};\\ &x_{01}=\left(\frac{K_1kT}{2\pi N_{01}e^2}\right)^{\frac{1}{2}};\quad x_{02}=\left(\frac{K_2kT}{2\pi N_{02}e^2}\right)^{\frac{1}{2}}.\end{aligned}\right\} \tag{4.1}$$

N_{01} and N_{02} are the densities at $x=0$, D respectively, and

$$\left.\begin{aligned} &\text{if } \alpha\leqslant d\quad 2K_2\alpha=K_1(D-d+x_{02})+K_2(d-x_{01}),\\ &\text{if } \alpha\geqslant d\quad 2K_1\alpha=K_1(D+d+x_{02})-K_2(d+x_{01}),\end{aligned}\right\} \tag{4.2}$$

so that if $K_1=K_2$ and $N_{01}=N_{02}$, the value of α is $D/2$. The condition determining which solution should be used is

$$\alpha\lessgtr d\quad\text{if}\quad K_1(D-d+x_{02})\lessgtr K_2(d+x_{01}). \tag{4.3}$$

As shown in the schematic diagram (Fig. 3), the potential jump at the boundary between the dielectrics depends only upon the dielectric constants and the difference of the metallic work functions, not upon the values of χ within the dielectrics.

The quantities x_0 are characteristic parameters emerging from the analysis. They have the dimensions of length, and may be regarded as scale factors affecting the mapping of the potential or field. A somewhat more intuitive picture may be obtained from the following considerations. Although the solution is not valid in the region beyond the dielectric, on the other hand, Eqs. (4) show that if it were, x_0 would be the distance into the metal at which the potential energy, field, and

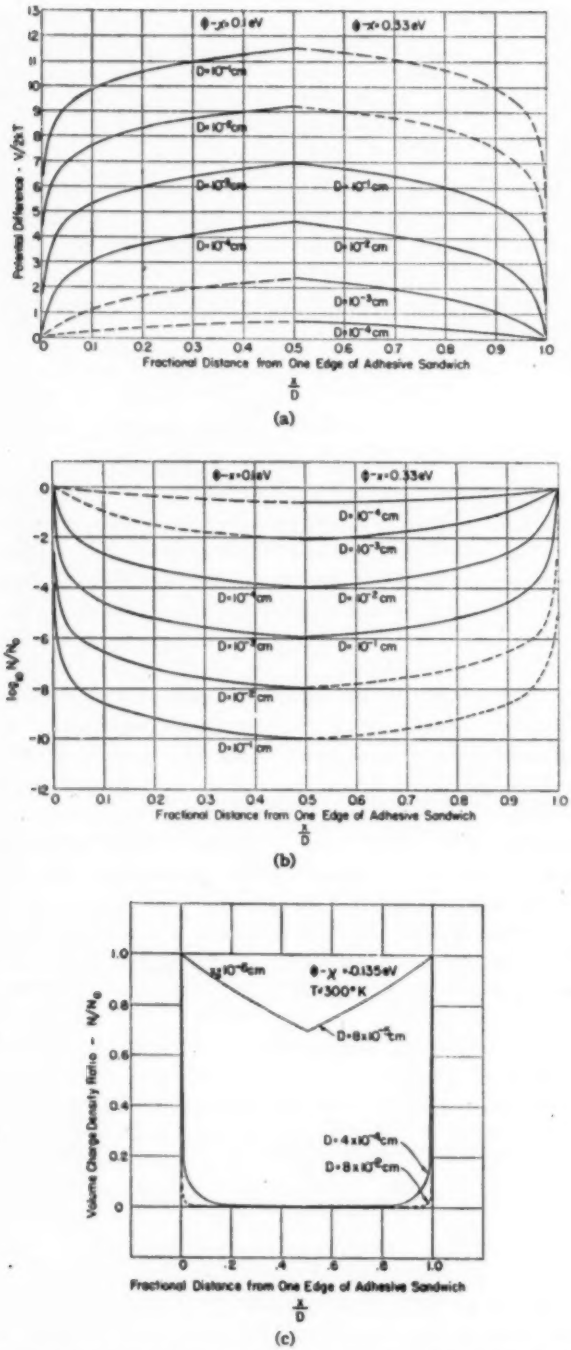


FIG. 4. (a) Potential in a bonded adhesive specimen. (b) Charge distribution in a bonded adhesive specimen. (c) The effect of the thickness of the dielectric layer on the charge distribution within the dielectric.

charge density become infinite; thus, the field (and other quantities) in the dielectric behave as if an infinite charge density existed at a distance (in nonexistent dielectric) of $-x_0$ from the boundary. Within the actually existing dielectric, the x_0 may be regarded as indicative of the range to which an appreciable influence of the metal penetrates [see Eq. (4.3)], recognizing of course that there is no sharp cutoff of any sort within any one dielectric. Obviously, the correct interpretation is as a scale factor; the dependence of this scale factor upon the energy levels in the dielectric and the metal is given by the definition in Eqs. (4) and by Eq. (5).

In order to determine the physical magnitude of the electron atmosphere in the dielectric, it is necessary to evaluate the N_0's. Obviously, these depend upon the nature of the metal and the dielectric in contact. Within the metal, the electron distribution is described by Fermi-Dirac statistics. At the boundary, the condition for equilibrium is that the absolute activities of the electrons in the dielectric and the metal shall be the same. This yields,[11] for barrier heights greater than approximately 0.1 ev at ordinary room temperatures

$$N_0=\frac{2(2\pi mkT)^{\frac{1}{2}}}{h^3}e^{-(\phi-\chi)/kT}. \tag{5}$$

To this approximation, the density corresponds to that portion of the high energy tail of the Fermi-Dirac distribution which exceeds the barrier energy. Table I gives values of N_0 and x_0 as a function of $\phi-\chi$. In certain cases, for better accuracy, the calculations of the above values have utilized interpolation between the values of the Fermi-Dirac integral tabulated by MacDougall and Stoner.[12] However, for very low barriers ($\phi-\chi=0.1$ or 0.05), Eq. (5) requires modification on statistical grounds. For the usual barriers, metal to air, the charge density is essentially zero, and x_0 is astronomical; (Table I, first line).

In usual thermionic measurements, only the saturation current and not N_0 can be measured; in the nonequilibrium separation (break) of an adhesive specimen (with a sufficiently rigid adhesive) the surface charge density (and thereby N_0) is measurable.

TABLE I. Boundary volume charge densities and values of the characteristic distance x_0 as a function of the barrier height and the dielectric constant. ($T=300°$.)

Barrier height $\phi-\chi$, ev	Volume charge density N_0 (cm^{-3})	x_0 (cm) for $K=1$	$K=4$	$K=10$
4.1	($<10^{-45}$)		($\cdots>10^{28}\cdots$)	
2.0	6×10^{-15}	2×10^{9}	4×10^{9}	7×10^{9}
1.5	2×10^{-6}	1.3×10^{5}	2.7×10^{5}	4.3×10^{5}
1.0	4×10^{3}	8.5×10^{0}	1.7×10^{1}	2.7×10^{1}
0.8	9.1×10^{5}	1.8×10^{-1}	3.5×10^{-1}	5.6×10^{-1}
0.6	2.1×10^{9}	3.7×10^{-3}	7.4×10^{-3}	1.2×10^{-2}
0.4	4.8×10^{12}	7.7×10^{-5}	1.5×10^{-4}	2.4×10^{-4}
0.3	2.3×10^{14}	1.1×10^{-5}	2.2×10^{-5}	3.5×10^{-5}
0.2	1.1×10^{16}	1.6×10^{-6}	3.2×10^{-6}	5.1×10^{-6}
0.1	4.9×10^{17}	2.4×10^{-7}	4.8×10^{-7}	7.6×10^{-7}
0.075	1.4×10^{18}	1.5×10^{-7}	2.9×10^{-7}	4.6×10^{-7}
0.05	3.5×10^{18}	0.9×10^{-7}	1.8×10^{-7}	2.9×10^{-7}

[11] R. H. Fowler and E. A. Guggenheim, reference 9, p. 475.
[12] J. MacDougall and E. C. Stoner, Trans. Roy. Soc. (London) A237, 67 (1938).

From the general case just derived, there may be obtained, by specialization, the two cases of interest, namely the equilibrium distribution of electron atmosphere in (1) a metal surface with adhering dielectric layer and another metal surface at a specified distance from the free adhesive side, hereafter called the dielectric film case; and (2) the bonded specimen. The former is obtained by letting $K_2=1$, and $D-d$ have any desired value. For the latter, $K_1=K_2$, and if the two plates are of the same metal, $N_{01}=N_{02}$. The result of letting $D=d$ in the dielectric film case is not at all the same as the case of the bonded sample. It corresponds rather to an unbonded nonadhesive contact, with an extremely small air film between the two. The dielectric film solution is also the equilibrium distribution, after break, and thus can be compared with the nonequilibrium existing immediately after break, corresponding to the bonded sample solution. It also represents the potential occurring at contact between a rigid nonmetallic solid and a metal.[13] The difference between the two cases is illustrated in Fig. 5. For practical purposes, the surface charge density on the metal in air is zero in the after break equilibrium picture; in a momentary contact, it would be a function of the time of contact, and the mobility and mean free path of the electrons in the dielectric. In the adhesive bond, on the contrary, there is an equilibrium charge density of a magnitude which is greater for smaller barrier heights.

V. THE SURFACE CHARGE

1. The Bonded Sample

If the same metal constitutes both plates, the sample is symmetrical with respect to its center plane, which, thermodynamically, could be replaced by a nonconducting boundary. The charge which has diffused into the adhesive from one plate is all contained in the nearer half of the adhesive.[14] The total charge per unit area of the metal surface will therefore be

$$\sigma=\int_0^{d/2} N(x)\cdot e\cdot dx=N_0 e\left(\frac{x_0 d/2}{x_0+d/2}\right).$$

If

$$x_0 \ll \frac{d}{2}, \quad \sigma=N_0 e x_0$$

and if

$$x_0 \gg \frac{d}{2}, \quad \sigma=N_0 e\left(\frac{d}{2}\right). \qquad (6)$$

If the specimen is broken in a time short enough so that the equilibrium charge distribution does not change appreciably, this expression will represent the surface density of charges remaining on the metal. Within the dielectric, the charge is distributed[15] as shown by N in Eq. (4). The rapid decrease of charge density from the interface into the interior is especially interesting. One of the consistent results which demanded explanation in the early experiments was the fact that a cohesive break (one occurring in the body of the adhesive rather than at the interface) showed little or no oscillograph trace with the voltage sensitivities used.[16] Figure 4c shows that for surface charge densities and adhesive thicknesses of the order of those measured, the volume charge density decreases by several orders of magnitude in the first 10^{-5} cm.

Figure 4 is drawn to show the potentials and charge densities for two different barrier values. Since in the bonded adhesive specimen with the same metal for both plates, these quantities are symmetrical about the center plane, one barrier value has been shown on the left and the other on the right. The mirror image curves

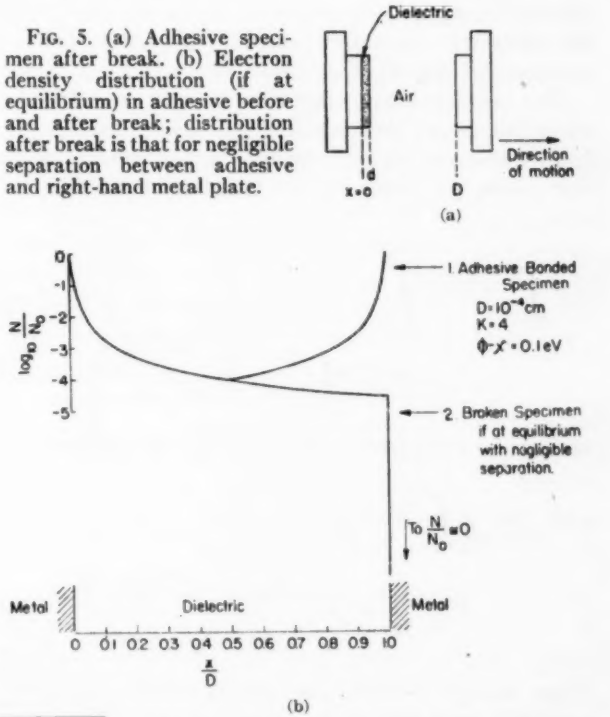


FIG. 5. (a) Adhesive specimen after break. (b) Electron density distribution (if at equilibrium) in adhesive before and after break; distribution after break is that for negligible separation between adhesive and right-hand metal plate.

[13] As, for example, in the measurements of R. M. Fuoss [J. Am. Chem. Soc. 59, 1703 (1937)], in which the equivalent of a layer of air of minimal thickness was corrected for.

[14] This is analogous to the classical picture of the lines of force from one metal plate ending on charges totally within that half of the adhesive nearest to it.

[15] The charge density distribution is sensitive to the exact model chosen. Surface states or acceptors and donors in the dielectric give a different volume or surface charge distribution in the dielectric. Therefore, experimental study of the distribution can contribute toward the choice between mechanisms. Qualitatively, the effects are sufficiently similar for our purpose.

[16] Actually, if there is a surface charge density in a cohesive break, a very much larger potential would be expected than for the same charge density in an adhesive break, as will be shown subsequently. Accordingly, the experimental results show that some charge, but of a much smaller order of magnitude, exists in a cohesive break.

(including the dotted ones) represent what would be the distributions in the other half, for the one barrier.

In any case, the accumulation of charge in the bonded adhesive must be a rate process, reaching eventual equilibrium during or after preparation of the sample.

For values of $\phi-\chi$ of the order of 0.1 ev, the concentration of the charge at equilibrium, in the layer near the surface in thick films permits consideration of the metal-polymer interface as a sharply defined double layer. However, the curves in Figs. 4 and 5 show that when the thickness of the adhesive film decreases, the charge spreads out through the volume, so that it must be treated as a diffuse double layer or a volume distribution.

2. After Break

In this case, the charge density at the boundary between the adhering film and the metal is N_{01}, as before, and the dielectric constant is the same. The broken surface, however, is in contact with air. In such case, it is necessary to use the solution which considers two different dielectrics. This solution differs considerably from that just discussed. For example, at the boundary between the adhering dielectric and the air, the charge density changes discontinuously in the inverse ratio of the dielectric constants, and the potential suffers a jump of $2kT\log(K_1/K_2)+(\phi_1-\phi_2)$.

The surface charge densities on the metal plates, when the second plate is still at essentially zero distance from the outer edge of the film which adheres to the first plate, are now: on the plate to which film still adheres

$$\sigma_1=\int_0^d Ne dx+\int_d^a Ne dx$$
$$=N_{01}e\left(\frac{x_{01}d}{d+x_{01}}\right)-\frac{K_1}{K_2}\frac{x_{02}}{d+x_{01}}\sigma_2, \quad (7.1)$$

and on the plate from which the film was stripped

$$\sigma_2=\lim_{(D-d)\to 0}\left[\int_D^a Ne dx\right]$$
$$=N_{02}e\frac{x_{02}[K_2(d+x_{01})-K_1x_{02}]}{K_1x_{02}+K_2(d+x_{01})}. \quad (7.2)$$

These charge distributions warrant some discussion, especially since this is the solution applicable both to the specimen after break (if it were permitted to come to equilibrium), and to the contact between a metal and a surface film on another metal. The plate on the right (Fig. 5) which, though stripped of film is still in virtual contact with the film, has a charge density which depends upon the sign of the factor in the numerator of Eq. (7.2); thus, if K_1x_{02} is more than $K_2(d+x_{01})$, this charge is of the same sign as the space charge in the dielectric. Since in practical cases of metal in contact with air, x_{02} is much larger than this value, it is evident that a very different physical situation exists here than existed in the bonded sandwich. *Where before, the second plate was attracted to the dielectric, now it is repelled by it.*[17]

The bonded joint before break is characterized by a space charge in the dielectric and an opposite charge on the metal. If the joint is broken, immediately the system is characterized by a new equilibrium condition toward which it tends, though it is far from equilibrium. Even while the separated plate is still effectively in contact with the dielectric, such equilibrium would consist of a reversal of charge so that the charge on the plate now would be of the same sign as that in the dielectric; the magnitude is (in practical cases) very much smaller than before. Similarly, if a dielectric film is already adhering to a metal plate, and another plate is brought up to the other side of the film, there will be a slight repulsion between the dielectric film and the new plate. Adhesion which is expected in a bonded joint, is not now to be expected, from the electrostatic forces arising from this (low temperature) thermionic mechanism. At large separations, it may be expected that the surface charge density on the second plate will not be affected by the conditions at the other metal plate and its adhering film. For copper in air, this is of course essentially zero.

In practice, other factors will enter. For example, the microscopic roughness and rigidity of solid materials will prevent close contact over a major fraction of the surface, so that the degree of electrostatic binding which is likely to exist will cause the metal to be charged in accordance with Eqs. (6) or (7), only in the vicinity of intimate contact between the two surfaces.

VI. THE ELECTROSTATIC CONTRIBUTION TO ADHESION

In addition to the London, Debye, and induction forces which have been utilized in the explanation of adhesion, it is evident that an electrostatic force of attraction exists between the volume charged dielectric and the oppositely charged surface. The energy supplied to break the adhesive sandwich rapidly must not only overcome any mechanical, van der Waals, dipole, etc. forces, but also produce separation of the charges. The existence of charges upon separation establishes this. What remains is to determine whether the electrostatic force is a significant contribution to the total force.

1. The Bonded Specimen

Although at atomic distances it is not justifiable to consider charges smoothed out into a uniform surface,

[17] Charge is transferred from the dielectric to the second plate (the one which has broken away), giving it charge of the same sign as the dielectric. If it were a neutral dielectric which furnished electrons to the metal, the approximation to the Fermi-Dirac integral, Eq. (5) (which is basic to Eq. (7.2)) would not be appropriate. However, the charge received by the second plate comes from the space charge in the dielectric, which originates in the electron sea of the first metal plate, and therefore Eq. (7.2) is valid.

the order of magnitude may be estimated in terms of such a picture. The classical electrostatic forces on the metal and dielectric include a tension on the metal directed toward the adhesive of magnitude

$f=2\pi\sigma^2/K$, which here, for the bonded film,

$$=\frac{2\pi N_{01}e^2}{K_1}\left[\frac{x_{01}(D/2)}{x_{01}+(D/2)}\right]^2=N_{01}kT\left[\frac{D/2}{x_{01}+(D/2)}\right]^2. \quad (8)$$

To a sufficient degree of accuracy, this is the electrostatic contribution to the force of adhesion.[18] As is true for the charge density, this force is dependent upon the ratio of the thickness of the film to x_0. When x_0 is much less than half the thickness of the adhesive film, the expression (8) becomes[19]

$$f=N_0kT \quad (8.1)$$

and if the adhesive film half-thickness is much less than x_0,

$$f=N_0kT\left(\frac{D}{2x_0}\right)^2=\frac{\pi e^2}{K}D^2N_0^2. \quad (8.2)$$

Accordingly, a dependence of strength upon thickness will be observed. The magnitude of the electrostatic contribution to adhesion increases rapidly with adhesive thickness at small thicknesses until $D/2$ is of the order of x_0, after which it approaches the constant value N_0kT, independent of how much thicker the adhesive layer is made. The greater the barrier height, the greater will be the critical adhesive film thickness, and for a given thickness, the smaller will be the electrostatic contribution to the adhesive force. Actually observed thickness-strength relationships show either increase, or decrease, of strength with decreasing thickness depending upon the adhesive and the metal.[20] It should be noted that for small barrier height (which is true in any case for which the electrostatic contribution is significant) the critical thickness is of the order of 10^{-6} or 10^{-7} cm, which is unlikely to be achieved in practice, so that a constant contribution is more likely to be observed.

[18] Within the adhesive the contained charge stresses the dielectric by the amount $(K/8\pi)F^2$. Thus the increase in adhesion is accompanied, classically, by a decrease in cohesive tensile breaking stress. That portion of the increased stress which is analogous to a uniform (negative) pressure must be added to any external tensile stress applied to the material. For example, if the dielectric is rigid enough to be treated by the Griffith or Sack crack theory, the term p_0 can be replaced by $(p_0+KF^2/8\pi)$ and the condition for spreading of the crack and eventual rupture of the material is then

$$p_0=\left(\frac{gE\gamma}{(1-\sigma^2)c}\right)^{\frac{1}{2}}-KF^2/8\pi$$

in which F is the field due to the contained charge, c is one-half the length or radius, respectively, of the crack, p_0 is the final "pressure" at which break occurs, σ, E, and γ, are the Poisson ratio, Young's modulus, and surface tension of the dielectric, respectively, and $g=2/\pi$ for the Griffith crack, and $\pi/2$ for the Sack (penny-shaped) crack, respectively.

[19] See also Fowler (reference 9), p. 366.

[20] G. W. Koehn, presented at the Symposium on Adhesion, Case Institute of Technology (April 24–25, 1952).

The variation of the electrostatic component of the adhesion force with temperature is determined by the two quantities N_0 and kT. A decrease of barrier $(\phi-\chi)$ increases N_0; temperature rise increases both N_0 and kT; thus, a fractional change in $(\phi-\chi)/kT$ by change of temperature produces more change in force than does the same fractional change produced by change in $\phi-\chi$ only. Of course, an increase of temperature produces other changes in the polymer which may override this. The temperature variation of f for large thickness is $f\propto T^{5/2}e^{-(\phi-\chi)/kT}$ and for small thickness or large barrier heights is $f\propto T^3e^{-2(\phi-\chi)/kT}$. Thus, any attempt to measure f with very thin films as a function of temperature would give a value of $\phi-\chi$ differing from that with thick films by a factor of two.

In practical cases, the microsurfaces of metals and polymers are generally rough and relatively rigid. Contact between them will occur at only a fraction of the total macroscopic contact area. Charges transferred at the points of contact will largely be bound into position electrostatically instead of being distributed evenly over the surface of the metal. Therefore, the charge is concentrated at points or small regions occupying only a fraction of the surface. Up to a certain point smoother surfaces should give greater electrostatic adhesion. The techniques used to produce suitable bonding, i.e., insuring wetting of the surface, or polymerizing in place from the monomer, will produce greater contact and therefore greater charge and consequent greater adhesion. Tacky surfaces and consequent flexibility at the interface will permit closer contact and the acquisition of the necessary charge over a larger fraction of the interface, thus promoting adhesion. Initial rigidity at the interface is prejudicial to and liquidity is favorable to adhesion. The same is true when other mechanisms are considered.[21]

2. The Sample after Break

For the cases dealt with here, namely metals bonded with polymers, the barrier to air renders x_{02} for the stripped metal so large that repulsion exists between the metal after break and the dielectric, if time is allowed for it to come to equilibrium. The actual break is so rapid that equilibrium is not achieved, and what is found is the charge distribution essentially characteristic of the bonded sandwich. Therefore, the same expressions as before (Eqs. (4) and (6)) may be accepted for the forces and charge densities.[22]

[21] For a discussion of the experimental data and review of such other mechanisms, see the succeeding article in this series.

[22] This is not true for the contact between an external metal plate brought up to an adhering film on another metal surface. When such a plate is brought up (in air) to the film, normally K_1x_{02} will be much more than $K_2(d+x_{01})$, so that there will exist a repulsion of approximately

$$\frac{2\pi\sigma_2^2}{K}=N_{02}kT\left[1-4\frac{K_2}{K_1}\frac{(d+x_{01})}{x_{02}}\right]. \quad (8.3)$$

With the values of N_{02} characteristic of even the metals with the lowest values of $\phi_{\text{thermionic}}$, this force is negligible; however, it is evident that electrostatically, adhesion is not favored.

TABLE II. Values of N_0kT ($T=300°K$).

N_0	N_0kT dyne cm^{-2}	N_0kT lb in^{-2}
10^{17}	4×10^{3}	0.06
(10^{19})	4×10^{5}	6
(10^{21})	4×10^{7}	600

However, it is of interest to examine what may occur in a very slow break. Assuming that contact is lost at a very small portion of the area by statistical fluctuations or otherwise, the charge in this region decreases, the electrostatic forces likewise decrease, and since x_{02} becomes very large in this small region, there is a tendency for like charge density to that in the dielectric to accumulate and consequently repulsion takes the place of adhesion. The repulsion increases the likelihood, under the existing tension, of neighboring areas being stripped, which then repeat the process. Accordingly, it is to be expected that a slow break will show less tensile strength, so far as the electrostatic contribution to adhesion is concerned, than will a fast break. Under practical conditions this effect may be overshadowed by other factors.

VII. THE EFFECT OF INTERMEDIATE ENERGY LEVELS

Actually the dielectric is more complex than the simple model of a region of uniform potential energy, $-\chi$. As in the Sommerfeld theory, the qualitative behavior is pretty well accounted for by the first approximation. However, when the various possible detailed structures are considered, there emerges a change in the volume electron distribution which is sensitive to the exact model.

If energy levels are available below the conduction band, the electron may enter such a level and temporarily be bound in its vicinity. Temperature agitation will eventually furnish sufficient energy, for the electron to enter the conduction band again. However, while it is in any such nonconducting level, the electron is contributing to the space charge between succeeding electrons and the interior of the dielectric. The net effect will be to cause a greater concentration of charge in the vicinity of the interface than is indicated by the previous treatment; this results in a greater field at the interface and consequent greater force. Similarly, available surface, or, in this case interface, states will cause an increase of charge density in the neighborhood of the interface.

The method of treatment of the first type of energy level is similar to the treatment of donors and acceptors given by Mott and Gurney,[9] and will not be sketched. For reasonable values of intermediate energy levels, the electrostatic forces at the interface are up to one order of magnitude greater than for the simple picture, and the charge distribution is intermediate between that of the simple model, and that of interface states.

An estimate of the effect of interface states may be made using the treatment of surface states by Bardeen,[22] developed for another purpose. Obviously, the effect sought is that corresponding to a considerable density of surface, or rather interface states, since the opposite extreme yields the model just discussed. Accordingly, his approximate solution, $l^2\approx l_0l_1$ is appropriate. The surface state density, alone, may be estimated from his Eq. (15), page 726,[22] yielding measurable values of the adhesive force as shown below. Taylor, Odell, and Fan[24] have shown that grain boundaries in germanium are characterized by surface states. While actual densities of such states are not known for metal polymer interfaces, it is reasonable to expect that such densities will be considerably higher than at interfaces between grains of approximately the same chemical constitution.

VIII. DISCUSSION

It remains to give some attention to the orders of magnitudes involved, and to various other possible explanations.

1. Orders of Magnitude

Table II gives some representative values of the electrostatic contribution to the adhesive breaking tension of bonded specimens for various barrier heights; practical adhesive film thicknesses in these cases will always be greater than x_0. Evidently on the simple model, which was used to discuss the observed charge densities occurring on break, the contribution to the adhesive tensile strength is only a small fraction of the total, except that it may be inferred that if values of N_0 of the order of 10^{21} cm^{-3} could be achieved, the contribution would be significant. In the current experiments, surface charges implying N_0 of the order of 10^{19} cm^{-3} have been obtained, with no definite experimental evidence of higher values, though these are not precluded.[6] Such a value corresponds to about one electron per 10^4 atoms in the dielectric and indicates barriers of 0.1 ev or less. Higher values would correspond to barrier heights of such sign and magnitude that an amplification of the present picture would be called for, for example, in the statistical mechanics of the equilibrium between the metal and dielectric at the interface.

With reasonable values for the magnitudes involved in the surface (interface) states, electrostatic forces at the interface up to 10^3 lb/in.^{2} may be computed, which is in the practical range.[25] Interface states alone do not

[22] J. Bardeen, Phys. Rev. 71, 717 (1947), appendix, pp. 726–7. See also F. Seitz (reference 9) pp. 320–6; J. J. Markham and P. H. Miller, Jr., Phys. Rev. 75, 959 (1949).

[24] Taylor, Odell, and Fan, Phys. Rev. 88, 867 (1952).

[25] If Bardeen's $(\epsilon_0-\varphi_0-\rho)/\epsilon_0$ is assumed to have the value 0.1, and n the value 10^{14}, $f=2\pi\sigma_s^2/K$ is of the order of 10^3 lb/in.^{2}. These are rather favorable values of the parameters; however, the contribution of the volume space charge has been neglected. Accordingly, it is reasonable to assume that interface levels or a combination of such levels with volume space charge can, under certain circumstances, result in adhesive forces of the order of 10^3 lb for 1 square inch specimens.

explain the charge densities observed at cohesive break, since they do not provide charge density in the interior of the dielectric.

2. Other Mechanisms

That values of $\phi - \chi$ in the range from 1 ev to 0.1 ev and even negative values exist in insulators is well known. That there exist charge densities in polymers bonded to metals is shown by the present experimental results. Various incidental references to what amounts to such charges may be found in the literature in fields other than adhesion.

An alternative explanation of the origin of the charges, is that they result from the rupture of dipoles which have been aligned at the interface due to the forces involved in adhesion. What is referred to here is not the surface electric moment due to the double layer, but the Debye dipoles resulting from the asymmetric distribution of charge within the polymer molecules. They exist throughout the volume of the polymer, are specific for particular chemical groups within the molecule, and are an accepted explanation of other polymer properties. However, while calculation of interactions between complete dipoles are relatively straightforward, quite arbitrary assumptions are necessary in the transition from dipole moments to charges resulting from their rupture. In addition, a thermodynamic treatment of such a mechanism meets the well-known difficulties arising in any attempt to combine electrical and mechanical quantities in a suitable thermodynamic cycle. Since charges are involved, it is more appropriate to consider the forces acting on such charges as a result of dipoles, image effects, etc., all of which are summed together in the determination of the value of χ. And, whether or not there are dipoles ruptured, the electron atmosphere or its equivalent will be present. Neither partial nor total dipole moments alone show any direct quantitative relation to cohesive forces,[26] so that it is somewhat optimistic to expect that they will do so to adhesive tensile strengths.

The possibility that the charge densities result from stripping a film of oxide from the surface of the metal has been considered. In such case, however, the charge densities found on the polymer faces in a cohesive break are difficult to understand unless a mechanism similar to that already treated is transferring electrons from the oxide into the polymer volume; this would qualitatively be similar to what has been discussed. The primary experimental fact would then be that the adhesion between polymer and oxide is greater than between oxide and metal. Experiments are under way to investigate what portion of the effects may be due to such an oxide stripping.

Certain other data in the literature bear on the present explanation, and at least raise the presumption that the flow of electrons into the polymer occurs in a manner analogous to thermionic emission. For example, Pohl[27] has injected electrons into rocksalt from a pointed electrode at 550°C; this cloud is visible and can be made to move back or forth by appropriate fields applied through plane parallel electrodes. It is known that color centers in alkali halides are related to electron traps, which may act as donors because of the small energy gap between their levels and the conduction band. Analogous motion of electrons in the interior of a dielectric occurs for semiconductors and crystal counters.

Meissner and Merrill[28] have reported that the adhesion of polystyrene to metal is small upon initial contact but increases with time of contact, and that the greater the pressure, the more rapid is the increase. It is well known that pressure sensitive adhesives show a gradual approach to final adhesive strength as a function of time. A portion of this time-dependent phenomenon would be the time required for establishment of the charge atmosphere in the available levels in the polystyrene. Such time dependence is inherent and natural in the electron atmosphere picture but must be brought in rather artificially in other mechanisms currently considered.

Havenhill, O'Brien, and Rankin[29] have found electrostatic contact potentials on various rubber and GR-S compounds after contact with clean metallic surfaces and later rapid (225 cm/sec) separation. The GR-S was negative with respect to the metal. By compounding it with electropositive materials the contact potential was reduced, and greater tensile strength in cohesive break was obtained. Some samples showed greater adhesion than cohesion, adhering to the metal plunger and breaking in the body of the material. As mentioned above, volume charges within the elastomer would produce an internal repulsion resulting in decrease in measured tensile strength for cohesive break. In another set of experiments, they related the charge picked up by a metal ball rolling down a rubber or GR-S sheet to the tensile strength of the material.[30] At high temperatures (see Eq. (5) above), the GR-S and rubber became more negative; the authors relate the greater decrease in tensile of GR-S over rubber to the greater "boiling off" of electrons. Milling less electronegative materials into the stock produced little effect unless a latex dispersion was used so that a relatively homo-

[26] N. A. deBruyne and R. Houwink, *Adhesion and Adhesives* (Elsevier Publishing Company, Inc., Houston, Texas, 1951), p. 16.

[27] R. W. Pohl, Physik Z. **35**, 107 (1934). See also N. Kalabuchow, "Transference of electrons from metals to dielectrics," Z. Physik **92**, 143 (1934).

[28] H. P. Meissner and E. W. Merrill, Am. Soc. Testing Materials Bull. **151**, 80 (1948).

[29] Havenhill, O'Brien, and Rankin, J. Appl. Phys. **15**, 731 (1944) and **17**, 338 (1946).

[30] This technique may be regarded as similar to the break of an adhesive sandwich, since there is a rapid separation of surfaces temporarily in contact. In the succeeding article, it will be shown that the measured potential in such an experiment is proportional to the charge density which has passed through or is on a square cm of the interface, i.e., σ in Eq. (6).

geneous material resulted, at which time noticeable effect was observed. Their Figs. 8, 10, and 11 are pertinent; in particular, their variation of the "contact potential" as a function of temperature agrees with Eq. (5) if the barrier is taken as 0.088 ev for the standard, and 0.069 ev for the silicated stock. The agreement is less good if a purely exponential function is assumed.

If the electrostatic contribution to the adhesive force is of some magnitude, the recent experiments of Meissner[31] are qualitatively explained by the above thermionic mechanism. Meissner measured the properties of unsupported thin films of various substances. His striking results included the fact that at and below approximately 0.25 micron thickness, the film of any substance measured adheres on contact with a clean metallic surface. After contact has been made to a metal adherend on one side, no adhesion is obtained upon bringing another clean metal surface up to the other side. Contact with the first metal surface would establish an atmosphere of electrons in the first portions of the film which touched, thus drawing up the remainder of the film to the metal (if its thickness were not so great as to require forces greater than those provided by the electron atmosphere), and eventually establishing the atmosphere in the whole film, with complete adherence. The situation would then be as in Fig. 5; air would exist between the adhering film and any second surface of metal approaching from the other side. Reference to Eq. (8.3), shows that such a second metal surface approaching the space charge filled adhesive film through the air would encounter a mild electrostatic repulsion, and adhesion would not result. Such phenomena would be true whether or not special techniques had been used to remove initial electrostatic charges from the films. If a barrier of not more than 0.3–0.4 ev is assumed, the critical thickness at which constant and therefore maximum (electrostatic) adhesive force is approached is 10^{-5} to 10^{-4} cm; thus, films of the thickness used by Meissner would have achieved their maximum electrostatic contribution to the adhesive force; the limitation of the universal adhesion to a thickness of 0.25 micron or less, of course, would be a matter of achieving a sufficiently thin film so that the force required to produce the deformation necessary for intimate contact between film and metal over microscopic irregularities would be no greater than that achievable by the term $f=N_0kT$.

[31] H. P. Meissner, reported at the Symposium on Adhesion held at Case Institute of Technology, April 24–25, 1952; published in J. Appl. Phys. **23**, 1170 (1952).

IX. CONCLUSION

If it is accepted that electrostatic charges received by the polymer from the metal by processes analogous to thermionic diffusion or its variants, and the resulting opposite charge on the metal, account for a portion of the adhesive forces, qualitative agreement with experiments is obtained, and explanation of certain previously unexplained phenomena results. The significance of the electrostatic contribution can only be determined by using the correct mechanism for the states into which the charge (positive or negative) goes in the polymer. One mechanism is treated in detail and others are briefly considered here. Insufficient information is at hand to determine whether the concentrations characteristic of the various mechanisms can reach magnitudes such that the electrostatic force becomes of an order of magnitude comparable with the total force necessary to produce break. If they can, the difficulty encountered in describing adhesive phenomena analytically using current mechanisms is understandable, since the electrostatic component depends upon factors such as the dielectric constant and the work function, factors which are not normally considered in fitting theory to experimental data when using traditional mechanisms. Lacking direct evidence, it is only possible to say that whatever effect the electrostatic terms contribute is qualitatively in detailed agreement with experiment in a number of different areas.

The whole phenomenon is complex. Many things are occurring at the same time. The electrostatic contribution is of course only a portion of the total forces acting; other forces operate, as previously considered in the literature, and calculations show that forces of the correct order of magnitude may also be obtained from them if suitable values for parameters are chosen.[32] Disentanglement of the various mechanisms and achievement of a means of estimating under what conditions each is important or dominant remains a formidable task for the experimenter.

[32] As, e.g., in S. J. Czyzak, Am. J. Phys. **20**, 440 (1952).

The Effect of Turbulence on Heat Transfer in a Pile with a Liquid Reactant

E. N. PARKER
University of Utah, Salt Lake City, Utah, and Unites States Naval Ordnance Test Station, Inyokern, California*
(Received December 3, 1952)

The spacing of cooling surfaces in a liquid reactant pile is considered in an elementary fashion. The resulting physical picture is that the eddy diffusion of thermal energy, which is enhanced by large free volumes of fluid, does not increase as fast as linearly with the spacing of cooling surfaces. Thus, for a given temperature difference between reactant and coolant, the total heat transfer decreases with increased spacing of the cooling surfaces. This disadvantage of large spacings is to be set off against the advantage of decreased parasitic neutron losses to the coolant.

INTRODUCTION

A PROBLEM of great importance in pile design is the removal of the thermal energy produced by the fission processes. Indeed, it is this heat transfer problem which is the main limiting factor in determining the minimum size of a pile for a given power output, provided, of course, that the pile is not so small that diffusion becomes a serious loss to the neutron flux within the pile.

With regard to the problem of heat transfer, it is of interest to consider the specific case of a pile using a liquid reactant. Presumably the heat exchange is carried out by circulating a coolant through tubes traversing the interior of the pile. Two extreme cases are possible: on the one hand we may increase the heat transfer by simply putting a very large amount of coolant tubing within the pile; on the other, we may hope to improve the transfer by reducing the amount of tubing, in which case the larger unimpeded volumes of reactant will then be free to develop convective currents and increase the heat transfer by the resulting turbulence. The latter alternative is very attractive because large amounts of tubing and coolant within the pile mean large parasitic losses to the neutron flux. In the following discussion we shall attempt to compare the heat transfer for the two cases. We shall establish an idealized flow pattern within a pile and calculate the variation of the heat transfer with the coolant tube spacing. Clearly our results will be valid for any heat transfer problem with the same boundary conditions.

PHYSICAL CONDITIONS

Consider a liquid reactant pile in which the coolant is piped through baffles, and assume that the baffles are held at some temperature T_0. The nature of the reactant, we assume, allows us to bring the temperature of the pile up to some value, say, T_1. To fix ideas as to baffle spacing, one possible sort of baffle arrangement is suggested in Fig. 1. We assume the baffles to be arranged to give maximum convection, and that only their separation is as yet undetermined. For a given T_0 and T_1 it is clear that the total convective force around the path of circulation $ABCD$ is essentially fixed. We wish now to compute the heat transfer as a function of the spacing between the baffles. With the above physical conditions in mind we solve the following idealized problem.

* Post Office: China Lake, California.

IDEALIZED MODEL

Let us consider the heat transfer from a two-dimensional flow between two infinite planes, $y=\pm a$. We assume the flow to be in the z direction and take the body forces producing the flow as independent of a. Thus, the Reynolds stress at the wall τ_0 is proportional[1] to a:

$$\tau_0 = C_1 a, \tag{1}$$

where C_1 is a proportionality constant. The velocity distribution near a plane wall is given[2] by

$$\frac{U}{U_\tau} = C + \frac{1}{K} \ln\left[\frac{(y \pm a) U_\tau}{\nu}\right], \tag{2}$$

where C and K are numerical constants and

$$U_\tau = \frac{\tau_0^{\frac{1}{2}}}{\rho} = \frac{(C_1 a)^{\frac{1}{2}}}{\rho}. \tag{3}$$

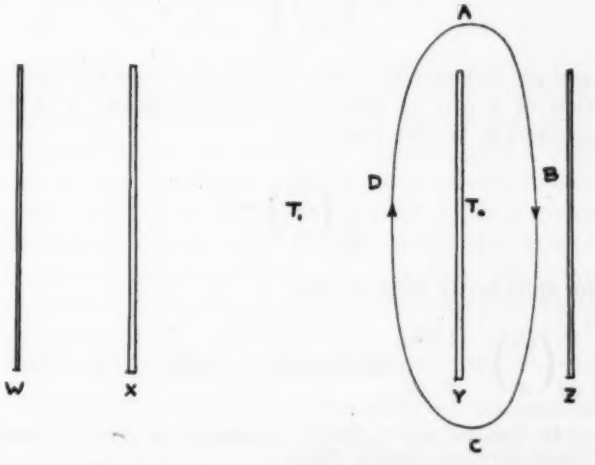


FIG. 1. Schematic view of hypothetical pile. W, X, Y, and Z are baffles containing coolant. D is the region of maximum heat generation.

[1] S. Goldstein, *Modern Developments in Fluid Dynamics* (Clarendon Press, Oxford, 1938), Eq. (1), p. 331.
[2] See reference 1, Eq. (8), p. 133.

Clearly Eq. (2) is not valid at the wall since the logarithm becomes infinite there. We see from Eq. (2) that the mean velocity of the fluid U_m is given roughly by an expression of the form

$$\frac{U_m}{U_\tau}\cong C_2 \ln\frac{aU_\tau}{\nu}+C_3, \tag{4}$$

where C_2 and C_3 are constants.

To determine the heat transfer, we use the expression[3]

$$Q_0=k_H\rho c U_m(T_m-T_0), \tag{5}$$

where T_m is the mean temperature across the channel, ρc is the specific heat per unit volume and k_H is the thermal transfer coefficient; Q_0 is the heat lost to the wall per unit time per unit area. We note that (T_m-T_0) and ρc are fixed so that we may write

$$Q_0=C_4 k_H U_m, \tag{6}$$

where C_4 is a constant. k_H is given in the literature[4] and is a function of a. The heat transfer per unit volume is thus

$$\frac{Q_0}{a}=C_4\frac{k_H U_m}{a}. \tag{7}$$

Substituting Eqs. (3) and (4) into (7) gives

$$\frac{Q_0}{a}=A\frac{k_H}{a^{\frac{1}{2}}}\ln a+B\frac{k_H}{a^{\frac{1}{2}}}, \tag{8}$$

where

$$A=\frac{3}{2}C_2C_4\left(\frac{C_1}{\rho}\right)^{\frac{1}{2}},$$

$$B=C_4\left(\frac{C_1}{\rho}\right)^{\frac{1}{2}}\left[C_2\ln\left(\frac{1}{\nu}\sqrt{\frac{C_1}{\rho}}\right)+C_3\right] \tag{9}$$

and are independent of a. k_H is a slowly varying function of a and is monotonically decreasing[5] with increasing a. We see then that

$$\frac{d}{da}\left(\frac{Q_0}{a}\right)<0$$

for sufficiently large a. Now

$$\frac{d}{da}\left(\frac{Q_0}{a}\right)=\frac{1}{a^{\frac{1}{2}}}\frac{dk_H}{da}(A\ln a+B)+\frac{k_H}{a^{\frac{1}{2}}}\left[A\left(1-\tfrac{1}{2}\ln a\right)-\tfrac{1}{2}B\right].$$

Thus, the maximum Q_0/a occurs where

$$0=\frac{1}{a^{\frac{1}{2}}}\frac{dk_H}{da}(A\ln a+B)+\frac{k_H}{a^{\frac{1}{2}}}\left[A\left(1-\tfrac{1}{2}\ln a\right)-\tfrac{1}{2}B\right]$$

or, denoting

$$\frac{a}{k_H}\frac{dk_H}{da}(A\ln a+B) \text{ by } -A\epsilon,$$

we have

$$\ln a=2-\frac{B}{A}-2\epsilon. \tag{10}$$

A, B, and ϵ are all greater than zero. From Eq. (9) we see that

$$\frac{B}{A}=\frac{2}{3}\left[\ln\left(\frac{1}{\nu}\frac{C_1^{\frac{1}{2}}}{\rho^{\frac{1}{2}}}\right)+\frac{C_3}{C_2}\right].$$

But aU_τ/ν is of the order of 10^4 or more. From Eq. (10) we see that $a<e^2$. Thus $(1/\nu)\,(C_1/\rho)^{\frac{1}{2}}$ must be of the order of 50 or more, and it follows at once that $B/A>2$. Hence $a<1$, which will not give turbulence for the values assumed. Or, in other words, if a is sufficiently large to allow turbulence, then

$$\frac{d}{da}\left(\frac{Q_0}{a}\right)<0.$$

We see then that by increasing a, we decrease the heat transfer per unit volume from the reactant.

Physically, we may understand the result as follows: The heat transfer per unit volume, neglecting the variation in k_H, is proportional to the mean flow velocity U_m divided by a. Thus, to gain in heat transfer, U_m must increase more rapidly than the first power of a. But, owing to the increased eddy viscosity if the channel is made wider, U_m does not increase this rapidly. Hence, Q_0/a decreases with increasing a, and increasing a is detrimental to the heat transfer problem. However, it should be noted that the neutron absorption of the baffles per joule of heat transferred is decreased since the heat transfer per unit area increases with a, i.e.,

$$\frac{dQ_0}{da}>0.$$

Quantitative results for the heat transfer with a specific geometry may be found in the literature,[6] though the computation is rather complex.

I should like to express my thanks to Dr. Lyle Borst who suggested the problem discussed above.

[3] H. Carslaw and J. Jaeger, *Conduction of Heat by Solids* (Clarendon Press, Oxford, 1947).
[4] See reference 1, p. 654.
[5] See reference 1, p. 660.
[6] Howard Schwartz, *Natural Convective Cooling of Liquid Homogeneous Reactors*, AECU-706, (NAA-*SR*-40).

A Retarding Potential Method for Measuring Electrical Conductivity of Oxide-Coated Cathodes*

I. L. SPARKS † AND H. R. PHILIPP
Department of Physics, University of Missouri, Columbia, Missouri
(Received November 10, 1952)

A retarding potential method is developed for measuring the electrical conductivity of normal oxide cathode coatings. The method is limited by normal current measuring devices and can not be used for coatings which have a conductivity to thermionic emission ratio greater than 2 cm/volt. Advantages of the method are: (1) the conductivity of coatings which are in a normal state for thermionic emission may be measured without the use of probe wires or other devices which might impair the thermionic emission of the sample, and (2) conductivity and thermionic emission measurements may be made simultaneously on the same coating sample. The theory of the method is discussed in detail and experimental results obtained using this method on both BaO and (BaSr)O coatings are given.

INTRODUCTION

FROM the viewpoint of theoretical considerations, simultaneous measurements of thermionic emission and electrical conductivity of oxide-coated cathodes are desirable. Investigators have encountered certain difficulties in the methods used for determining the electrical conductivity. This has been especially true when attempts were made to study thermionic emission from the same sample. The method[1] of embedding a small probe wire in the coating has probably been used more widely than other methods. This is usually accomplished by spraying a small amount of coating on a cylindrical base metal, winding a fine platinum wire around this, and again spraying until the wire is completely covered. In general, a nonuniform coating surface has resulted owing to the irregularity introduced by the probe wire. Another disadvantage to this method is that an unusually thick coating is required to completely cover the wire. Abnormally thick coatings often give a low thermionic emission and are prone to develop cracks and adhere poorly to the base metal.

Hannay, MacNair, and White[2] used a ceramic base cathode sleeve around which a pair of conductivity leads were wound. The oxide coating was sprayed over this base. Although both conductivity and thermionic emission were readily measured, it was not possible to extend these studies below the temperature range in which an optical pyrometer could be used. The absence of a continuous base metal prevented the use of thermocouples for a temperature determination.

Loosjes and Vink's[3] method of pressing the coating between two flat base metals was satisfactory for determining the conductivity but did not allow a convenient means of measuring the thermionic emission. In order to determine the thermionic emission, it was necessary to separate the two cathodes and insert an anode between them after the conductivity measurements were made. This, of course, involves the hazard of damaging the coating and of changing surface activity as was indeed observed by them. A similar method was used by Young[4] in which the thermionic emission was taken to a ring anode. Accurate measurements of thermionic emission were not possible as the emitting area was not well defined.

With these difficulties in mind, it was thought desirable to attempt to devise a method of measuring the conductivity which would leave the coating in its normal state for thermionic emission. As thermionic emission current is drawn from a cathode, a voltage drop develops across the coating. This arises from the fact that the coating has a finite conductivity and a current equal to the emission current passes through it. If this voltage drop can be measured as a function of the emission current, the resistance of the coating can be determined. A knowledge of the geometry of the coating will enable one to calculate the specific conductivity.

Young and Eisenstein[5] have reported on a retarding potential method which would enable one to determine the voltage drop across the coating. In this work they assumed that the Schottky[6] lowering of the potential barrier at the surface of the cathode was negligible compared to the voltage drop across the coating. They also assumed that the method was applicable for a wide range of anode voltages. In the present investigation a similar method is developed without making these assumptions. Young and Eisenstein found the slope of the reatarding potential curve was less than the expected value, $-e/kT$. A possible explanation of this is given here.

EXPERIMENTAL TUBE

A schematic drawing of the experimental tube used in this investigation is shown in Fig. 1. The indirectly heated cathode, the type used in the 2C39 lighthouse

* Supported in part by the U. S. Office of Naval Research.
† Now at Eastern Illinois State College, Charleston, Illinois.
[1] G. W. Mahlman, J. Appl. Phys. **20**, 197 (1949).
[2] Hannay, MacNair, and White, J. Appl. Phys. **20**, 669 (1946).
[3] R. Loosjes and H. J. Vink, Philips Res. Rep. **4**, 449 (1949).
[4] J. R. Young, J. Appl. Phys. **23**, 1129 (1952).
[5] J. R. Young and A. S. Eisenstein, Phys. Rev. **75**, 347(A) (1949).
[6] F. Seitz, *The Modern Theory of Solids* (McGraw-Hill Book Company, Inc., New York, 1940), pp. 162.

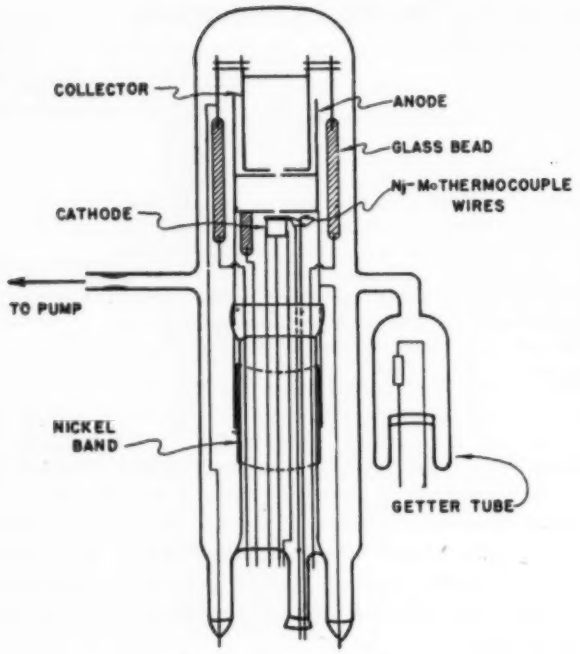


FIG. 1. Experimental tube.

tube, has a flat 0.5 cm² pure nickel‡ button welded to the top. A Ni-Mo thermocouple is used to determine the temperature of the cathode base metal. Electrons passing through the pair of one-mm holes, 10 mm apart in the tantalum anode, enter the tantalum collector through a 6-mm hole. The collector is supported by Nonex beads and the anode by a nickel band around the main press. In order to minimize electrical leakage, electrical connections to the anode and collector enter the tube through separate extended presses. A batalum getter, shown in the small attached side tube, was flashed immediately after seal-off. Figure 2 shows the electrometer tube circuit with the G.E. Type 5674 electrometer tube, which is used to measure the collector current. This circuit is capable of measuring collector currents in the range 10^{-9} to 10^{-15} ampere. Collector voltages are measured with a 0–16 volt range Leeds and Northrup Type K-2 potentiometer.

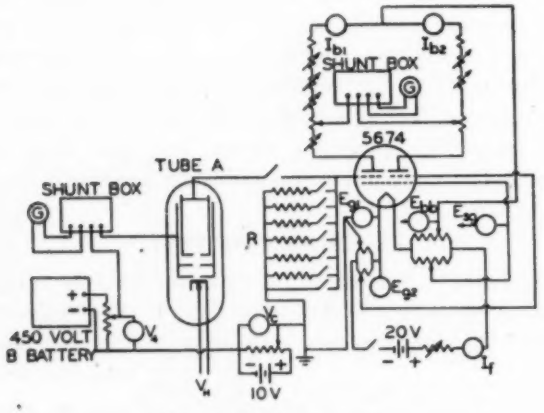


FIG. 2. Circuit diagram.

‡ 1001 electrolytic nickel obtained from E. M. Wise, International Nickel Company.

THEORY OF THE RETARDING POTENTIAL METHOD

An energy level diagram for the cathode, anode, and collector of the experimental tube is shown in Fig. 3. The diagram has been simplified somewhat by the omission of impurity levels of the oxide and the levels of the filled bands, since they are not necessary for this discussion. Diagrams A, B, and C represent the situations for different values of anode and collector potentials. In A, the anode and collector are at cathode potential. In this case, the Fermi level of the base metal, the

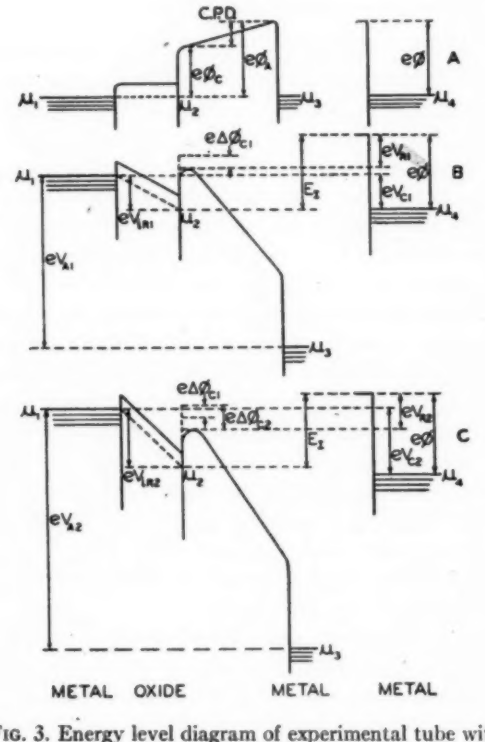


FIG. 3. Energy level diagram of experimental tube with anode at different potentials.

chemical potential of the oxide coating, and the Fermi levels of the anode and collector are at the same energy These are designated by the symbols μ_1, μ_2, μ_3, μ_4, respectively. If the work function of the anode, ϕ_A, is greater than the work function of the cathode, ϕ_C, a retarding contact potential difference (C.P.D.) equal to $(\phi_A - \phi_C)$ will appear between the two surfaces. Since the anode and collector of the experimental tube were both tantalum, the work functions of these are shown to be about equal. The application of a large accelerating potential, V_{A1}, to the anode displaces the Fermi level of the anode μ_3 downward by this amount as shown in *B* and permits a flow of electron current from the cathode to the anode. The presence of a high electric field at the cathode causes a Schottky lowering of

its surface barrier by an amount designated as $e\Delta\phi_{C_1}$. Due to the finite conductivity of the oxide, a voltage drop V_{iR_1}, will develop causing a tipping of the energy bands in the oxide. Electrons leaving the cathode will arrive at the anode with a minimum kinetic energy of

$$K.E._{\min} = eV_{A_1} - eV_{iR_1} - e\Delta\phi_{C_1} - e(\text{C.P.D.}), \quad (1)$$

since this is the kinetic energy gained in passing between the two surfaces. Of course, most electrons will have energies greater than this owing to the energy distribution of the electrons leaving the surface of the cathode. A fraction of the electrons arriving at the anode will pass through the one mm hole and reach the collector with kinetic energies which depend on the collector potential V_C. If the potential of the collector is positive with respect to the cathode and greater than $(\phi_C - \Delta\phi_{C_1})$, (i.e., the surface barrier of the collector is below the surface barrier of the cathode in the diagram), all electrons passing through the hole in the anode will reach the collector. As the collector potential is made less and less positive, the current reaching the collector remains constant until the surface barrier of the collector is at the same level as the cathode barrier. Beyond this point, the collector surface barrier becomes negative with respect to that of the cathode and electrons arriving at the anode with the minimum kinetic energy will be repelled by the collector. As the collector is made more negative, only the higher energy electrons will be collected. Thus, if one makes a plot of log collector current, I_C, *versus* collector voltage, V_C, the curve should show a break at a collector voltage for which the surface barriers of the cathode and collector are at the same level, Fig. 4. This would occur at a point where V_R, the true cathode-collector retarding potential, has the value zero.

If now an accelerating voltage, V_{A_2}, $(V_{A_2} > V_{A_1})$, is placed on the anode, the Fermi level μ_3 will be displaced downward, Fig. 3-C. The surface barrier of the cathode

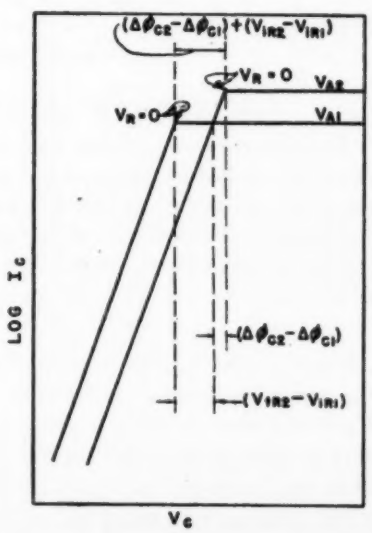


FIG. 4. Retarding potential curves expected from discussion of Fig. 3.

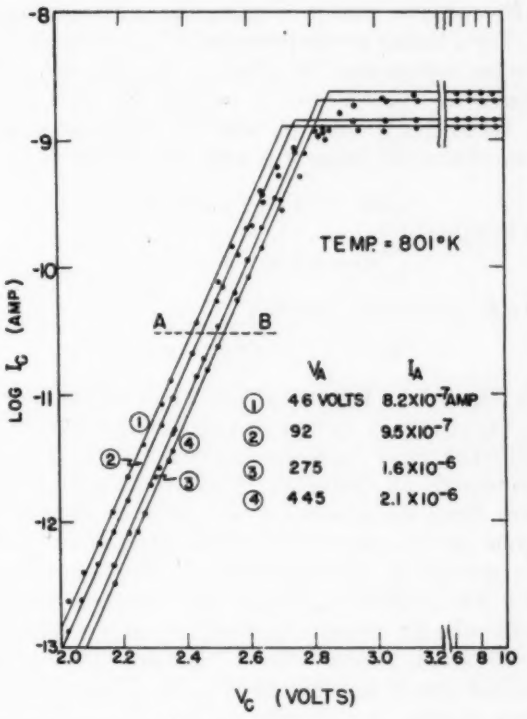


FIG. 5. Experimentally determined retarding potential curves for different anode potentials.

will be lowered by an amount $e\Delta\phi_{C_2}$, $(\Delta\phi_{C_2} > \Delta\phi_{C_1})$ increasing the anode current.§ A larger voltage drop across the coating, V_{iR_2}, will occur due to this increased anode current. A new plot of $\log I_C$ *versus* V_C will yield a curve similar to the previous one except the saturated collector current will now have a higher value due to the increase in anode current, Fig. 4. Also, the break $(V_R = 0)$ in the curve should occur for a value of V_C somewhat higher than before since the surface barrier of the cathode is now lower than that in case B by an amount ΔE,

$$E = e(\Delta\phi_{C_2} - \Delta\phi_{C_1}) + e(V_{iR_2} - V_{iR_1}). \quad (2)$$

COMPARISON WITH EXPERIMENT

Four retarding potential curves obtained experimentally in the manner described above for different values of V_A are shown in Fig. 5. The shift of the curves to the right with increasing current can be seen. The scale is broken at $V_C = 3.2$ volts in order to show that the saturation collector current remains constant to 10 volts

If the collector voltage, V_C, is adjusted to a particular value V_{C_1}, Fig. 3-B, such that V_{R_1} will have a negative value, the energy difference between the collector vacuum surface level and μ_2 at the cathode surface will

§ In studies of this kind, it is usually assumed that the transmission coefficient for electrons at the surface barrier of the cathode is equal to unity and does not change appreciably with V_A over the voltage range considered here. It is further assumed that the energy distribution of emitted electrons does not change with respect to μ_2 at the surface.

be E_I and give rise to a particular collector current, Ic_1. For a higher anode potential, V_{A_2}, in Fig. 3-C, the collector voltage must be adjusted to a new value, Vc_2, in order for the collector to receive the same current Ic_1. When this is done, E_I will have the same value as before. From the figure it is seen that in case B,

$$E_I = e(V_{iR_1} - Vc_1 + \phi) \tag{3}$$

and in case C,

$$E_I = e(V_{iR_1} - Vc_2 + \phi). \tag{4}$$

Since E_I and ϕ are constant,

$$(Vc_2 - Vc_1) = (V_{iR_2} - V_{iR_1}). \tag{5}$$

In the work by Young and Eisenstein,[5] it was assumed that the change in voltage drop across the coating, due to different anode currents, was equal to the difference in potentials at the break in the retarding potential curves. Since the presence of any space charge near the cathode surface affects the location of the break point, their method is not applicable in the space charge region. This limitation has been foreseen by Wright and Woods[7] who use a similar method but make a correction for the space charge that is present. The method described here is not limited by space charge since the surface barrier of the collector is at a much higher potential than the space charge barrier at the operating point where conductivity determinations are made. In the work mentioned above, it was also assumed that the Schottky lowering of the surface barrier of the cathode was negligible compared with the voltage drop across the coating. From the discussion in the previous section and the method of taking data described below, it is apparent that the present method is not limited by the space charge barrier and the Schottky lowering of the barrier at the surface of the cathode is not neglected.

The horizontal line AB in Fig. 5 represents a typical constant collector current value, I_C. From Eq. (5) the change in V_C is equal to the change of voltage drop across the coating corresponding to four values of current I_A through the coating. A plot of V_C *versus* I_A should yield a straight line, the slope of which is equal to the effective resistance of the coating, R. A knowledge of the coating thickness t, and area A, gives the effective specific conductivity.

$$\sigma = t/(RA). \tag{6}$$

Since complete retarding potential curves of the type shown in Fig. 5 required considerable time to obtain, a faster method was devised to determine the conductivity. At a given temperature and at a particular anode current, the saturated collector current was measured by setting the collector at +10 volts. Then a collector current, one or two orders of magnitude lower than the saturation current was selected. As the anode current was varied by changing the anode voltage, the collector voltage necessary to give the selected collector current was measured. This effectively determines the points of intersection of the line AB with the retarding potential curves, Fig. 5. In this manner, a series of these points could be determined in a relatively short time.

A typical plot of these intersection voltages, V_C *vs* I_A used to evaluate σ is shown in Fig. 6. The anode voltage range covered to obtain the anode currents for this curve was 46 to 455 volts. The experimental points fall close to a straight line in the high voltage range but deviate somewhat at the three lower currents, a situation found for all curves of this type. The theory discussed above assumed that the current reaching the collector was a definite fraction of the anode current, therefore, a constant ratio of anode current to saturated collector current should be obtained. Table I shows the measured ratio of anode current, I_A, to saturated collector current, Ic_s, for the anode voltages used in obtaining Fig. 6. It is seen that the ratio decreases for the three points falling below the straight line in Fig. 6 thus explaining the reason for their deviation. The exact cause of this effect is not known, but in the subsequent measurement of conductivity, only anode potentials above 180 volts were used to avoid this difficulty. From the straight line section of this curve a conductivity of $5.7\times10^{-7}\Omega^{-1}$ cm^{-1} is computed using Eq. (6).

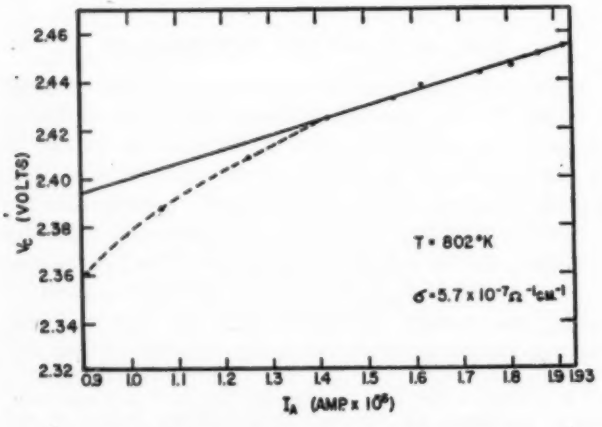


FIG. 6. Collector voltage *versus* anode current.

ELECTRON ENERGY DISTRIBUTION

For the case of a diode with concentric cylindrical geometry, Schottky[8] derived an equation for thermionic emission in retarding fields, assuming a Maxwell-Boltzmann distribution of electron velocities. This equation reduces to a simple form for the plane parallel geometry case:

$$J_R = J_0 e^{-eV_R/kT} \tag{7}$$

where J_R is the current density reaching the anode, J_0 is the zero field thermionic emission current density, and V_R is the true retarding potential between the surface of the cathode and the surface of the anode. If one plots log J_R *versus* V_R, a straight line should result, the

[7] D. A. Wright and J. Woods, Proc. Phys. Soc. (London), B, 65, 134 (1952).

[8] W. Schottky, Ann. Physik 44, 1011 (1914).

slope of which is $-e/kT$. Hung[9] investigated the emission from oxide-coated cathodes in retarding fields and found that in the high energy region the experimental curve was a straight line and the temperature calculated from the slope agreed well with the observed temperature. Some deviation from the theoretical curve was found in the low energy region which he attributed to space-charge effects.

In the case of the plane parallel geometry diode, it is apparent that only the normal component of the electron velocity determines whether a given electron will reach the anode. It is also apparent that because of the complicated field arrangement in the experimental tube used in this investigation, it would be quite difficult to calculate accurately the energy distribution of electrons arriving at the collector. It is interesting to note, however, that, if one assumes that the total initial velocity, rather than the normal component of velocity, determines whether an electron will reach the collector, a somewhat different expression is obtained for the collector current and with high anode potentials this will probably be the case. An electron emitted from the cathode has both normal and parallel velocity components but under the influence of the accelerating field will arrive at the anode moving in essentially a normal direction. Since the normal component of velocity is large compared with the parallel components, any electron passing through the anode will do so with negligible parallel displacement. It is only as the electron approaches the collector and is retarded that the initial velocity components again influence the direction of motion. Since the collector is a hollow cylinder located near the anode exit aperture all electrons passing through the anode will be collected regardless of their parallel velocity components. Assuming that the electrons in the oxide with suffiicent energy to escape have a Maxwell-Boltzmann distribution, the number per unit volume having velocity components in the range, dv_x, dv_y, and dv_z is given by

$$N(v_x, v_y, v_z)dv_x dv_y dv_z = n(2\pi kT/m)^{-\frac{1}{2}} \exp-[mv^2/2kT]dv_x dv_y dv_z \quad (8)$$

where n is the number per unit volume with all velocities. The velocity distribution of those escaping unit area per unit time will be this number multiplied by v_x with x taken as the direction normal to the surface. To determine the number leaving unit area per unit time having a speed between v and $v+dv$, the number leaving per unit time must be integrated over a semispherical shell of thickness dv. That is,

$$\int\int\int v_x N(v_x, v_y, v_z)dv_x dv_y dv_z = \int_v^{v+dv}\int_0^{\pi/2}\int_0^{2\pi} v^3 N(v_x, v_y, v_z)\sin\theta\cos\theta dv d\theta d\phi \quad (9)$$

[9] C. S. Hung, J. Appl. Phys. **21**, 37 (1950).

TABLE I. Anode current-saturated collector current ratios.

V_A	I_A/I_C
455	823
410	817
365	830
320	837
275	832
227	828
183	815
137	723
92	644
46	715

where $v^2\sin\theta dv d\theta d\phi$ is the volume element in v space, and $v_x = v\cos\theta$. Carrying out the integration and expressing the result in terms of energy, the number of electrons escaping unit area per unit time with energy between E and $E+dE$ is given by

$$g(E)dE = CEe^{-E/kT}dE \quad (10)$$

where $C = n(2\pi m)^{-\frac{1}{2}}(kT)^{-\frac{1}{2}}$. The number of electrons arriving at the collector will be the fraction, f, of these which pass through the anode aperture with energies greater than eV_R and the current will be the electronic charge multiplied by this number.

$$I_C = efC\int_{eV_R}^{\infty} Ee^{-E/kT}dE = efn(2\pi mkT)^{-\frac{1}{2}}(kT+eV_R)e^{-eV_R/kT}. \quad (11)$$

For a given temperature this may be expressed as,

$$I_C = K(kT+eV_R)e^{-eV_R/kT}. \quad (12)$$

According to this equation, a plot of $\log I_C/(kT+eV_R)$ *vs* V_R should yield a straight line with slope $-e/kT$.

Figure 7 shows $\log I_C$ *vs* V_C for two different temperatures, as would be plotted for the usual retarding potential case. Plotting V_C, the applied voltage, rather than V_R merely shifts the curve to the right by the contact potential difference. These curves differ from the usual diode retarding curves in three respects, (1) the saturated portion is much flatter than is usually obtained, (2) the break is sharper, and (3) the temperature, calculated by assuming the slope equal to $-e/kT$, is in poor agreement with the measured temperature as shown on the figure.

Figure 8 shows the same data plotted as $\log I_C/(kT+eV_R)$ *versus* V_R. The true retarding voltage V_R was determined by assuming $V_R=0$ at the breakpoints in Fig. 7. The temperature was calculated again assuming the slope equal to $-e/kT$ and was found to agree closely with the measured temperature as indicated on the figure.

Thus, the experimental observations seem to favor the latter expression for the collector current. Whether this is the exact expression or not is not important as far as the theory of conductivity measurement is con-

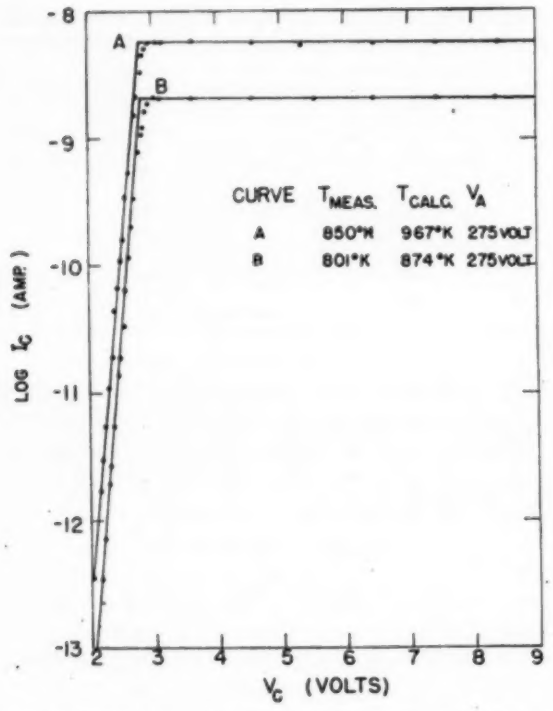


FIG. 7. Retarding potential curves as would be plotted for diode with plane parallel geometry.

cerned. In this theory the only assumption concerning the energy distribution is that the electron density at some level above the cathode surface level remains constant with respect to μ_2 as V_A is changed. This seems to be a valid assumption since the retarding potential curves at a given temperature, Fig. 5, are parallel in the high energy region.

APPLICATION OF THE RETARDING POTENTIAL METHOD

Results obtained in applying this method to five different cathodes will be given. For one cathode, results will be given with the cathode in three different states of activation. These results are summarized in Table II. Carbonates used in spraying the cathodes were equal molar $(BaSr)CO_3$[‖] and $BaCO_3$.[¶] The cathodes had a coating weight of approximately 10 mg/cm², the density approximately 1 g/cm³ and the area was 0.5 cm².

Conductivity data at various temperatures were obtained for all cathodes except VI and VII. Three typical curves of collector voltage, V_C, *versus* anode current, I_A, which were obtained for cathode II, are given in Fig. 9. The resistance of the coating at these temperatures was obtained from the slopes of the curves. Coating conductivity values were calculated using Eq. (6) and will be presented later.

An attempt was made to measure the coating conductivity for cathodes VI and VII using this method. Over an approximate temperature range 550 to 1000°K the change in voltage drop across the coating ΔV, for these two cathodes, was too small to be measured. The stability of the experimental apparatus was such that a ΔV of 0.005 volt could have been detected. Since the coating voltage increases with increasing thermionic emission J_0, and decreases with increasing conductivity σ, it is apparent that the ratio, σ/J_0, of these two cathodes was too large (i.e., giving a voltage drop too small to be detected) in order for the conductivity to be measured by this method. This ratio σ/J_0 then permits us to calculate the range over which this method can be used for determining the coating conductivity.

The importance attached to the ratio σ/J_0 in setting the range over which this conductivity method can be used is seen by a simple calculation. Taking the minimum measurable voltage change to represent essentially the total coating voltage, as an upper limit calculation, this voltage may be expressed as $V=J_0t/\sigma$, where t is the coating thickness and the ratio $\sigma/J_0=t/V$. Setting $V=0.005$ volt and $t=0.01$ cm gives $\sigma/J_0=2$ cm/volt. Thus, cathodes for which $\sigma/J_0>2$ cannot be expected to have a coating voltage greater than 0.005 volt.

Data on both the conductivity and thermionic emission of (BaSr)O over the same temperature range are found in the literature in only a few cases. Hannay,

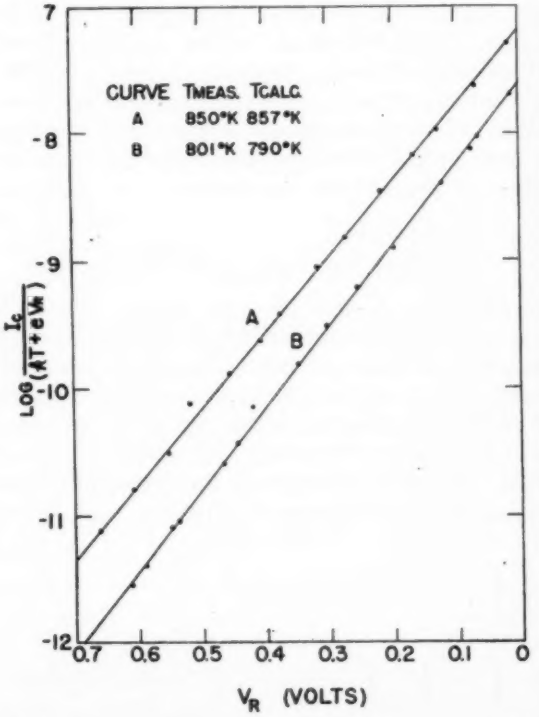


FIG. 8. Retarding potential curves plotted for experimental tube.

‖ C51-2 obtained from Raytheon Manufacturing Company.

¶ Ultra-Pure $BaCO_3$ from Mallinckrodt Chemical Works (30 parts per million of Sr) used on cathode V and a carbonate prepared in this laboratory by a process involving the recrystallization of $Ba(NO_3)_2$ (less than 1 part per million of Sr) used on cathodes VI and VII. Preparation carried out by Mr. Harold John.

TABLE II. Summary of results.

Cathode number	Material	$e\phi$ (eV)	Q_2 (eV)	Q_1 (eV)	J_0(1000°K) (amp/cm²)	σ(1000°K) (ohm⁻¹ cm⁻¹)	σ/J_0(1000°K) (cm/volt)	Break-point in conductivity curve temp. °K
I	(BaSr)O	1.25	1.18	0.39	2.00×10^{-2}	1.55×10^{-3}	7.75×10^{-2}	720
II	(BaSr)O	1.42	0.92	0.41	1.10×10^{-2}	1.35×10^{-3}	1.23×10^{-1}	748
III	(BaSr)O	1.38	1.04	0.41	7.50×10^{-3}	5.00×10^{-4}	6.67×10^{-2}	742
IV	(BaSr)O	1.34	1.42		1.26×10^{-2}	1.60×10^{-3}	1.27×10^{-1}	<700
V	BaO	1.79	1.59	0.62	1.80×10^{-4}	5.23×10^{-5}	2.90×10^{-1}	750
VI	BaO	1.37			1.36×10^{-2}		>2	
VII	BaO	1.50			1.29×10^{-3}		>2	

MacNair, and White[2] show data from which the ratio σ/J_0 can be calculated. Their value is approxiamtely 0.03 at 1000°K. Mahlman's[1] data give a ratio of 0.04 at 770°K and the value for (BaSr)O reported in Table II, in each case, is less than 2 at 1000°K. The value for cathode V (BaO) is also less than 2, while for VI and VII it is evidently greater than 2. Thus, it is seen that the method is applicable for the (BaSr)O cathodes which are reported but is applicable only in certain cases for BaO.

CONDUCTIVITY AND THERMIONIC EMISSION AS A FUNCTION OF TEMPERATURE

The experimentally observed variation of the oxide coating conductivity with temperature is usually expressed[10]

$$\sigma = Ke^{-Q/kT}. \tag{13}$$

Over the temperature range form 300°K to approximately 1000°K it has been found that the plot of log σ *versus* $1/T$ may show two linear regions with appropriate values of K and Q for each region. Loosjes and Vink[3] attribute their results of this type as being due to two different conduction mechanisms in parallel. In region I (the low temperature region), the measured conductivity is predominately due to electron conduction through the crystals of the coating. The electrons pass from one crystal to the next at the point of contact. In region II (the high temperature region), the observed conductivity arises primarily from a conduction by the electron gas in the pores between the crystals. Electrons thermionically emitted from the crystals form the electron gas in the pores and may then pass from pore to pore. A more recent explanation[11] supposes that the surface and bulk conductivity of the crystals are different; thus two parallel conduction paths result.

The thermionic emission of an oxide coated cathode is usually shown by a Richardson plot in which the logarithm of the zero field emission current density divided by the square of the absolute temperature, $\log J_0/T^2$, is plotted as a function of reciprocal temperature, $1/T$. For all cathodes studied this plot gave a straight line. The slope of the Richardson plot is equal to $-e\phi/K$, where k is Boltzmann's constant. From the slope, the apparent work function, $e\phi$, was determined.

After the cathodes were processed and aged by drawing a small emission current from the cathode for 24 hours, conductivity and thermionic emission measurements were made as a function of temperature. Figure 10 shows the temperature dependence of the conductivity for cathode V. A typical curve for the (BaSr)O cathodes (cathode II) is shown in Fig. 11. These curves are similar to the type observed by Young[4] and also by Loosjes and Vink.[3] The conductivity of cathode IV was measured over the range 700 to 1000°K. Since the plot of log σ *versus* $1/T$ for this cathode was a straight line, the break in the curve evidently occurs below 700°K. Using a different method of measuring conductivity others[3, 4] have found similar straight line conductivity plots, particularly with cathodes in a low state of activation.

The zero field emission current density, J_0, was deter-

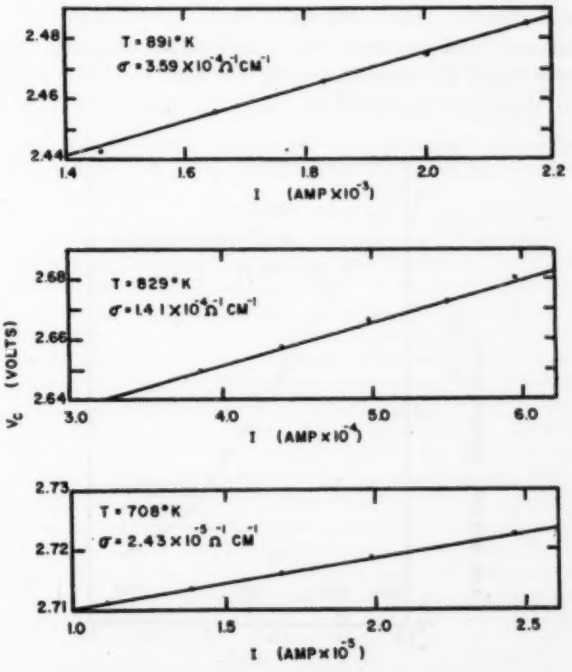


FIG. 9. Coating voltage as a function of coating current for cathode II.

[10] A. Eisenstein, *Advances in Electronics* (Academic Press, Inc., New York, 1948) Vol. 1, Pt. 1.

[11] D. A. Wright, paper submitted to the Symposium on Electron Emission, (New York City), January 30, 1951.

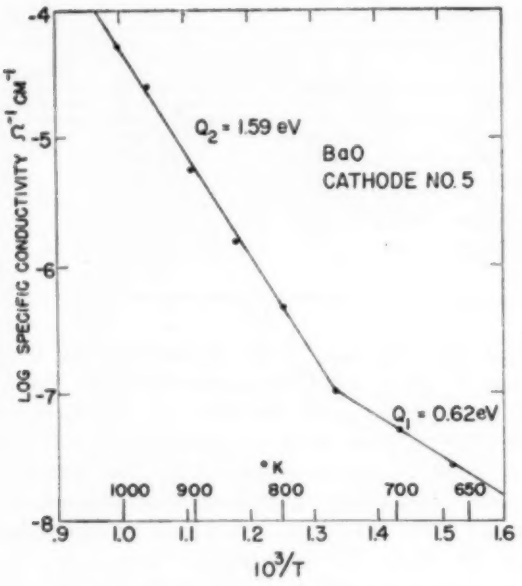


FIG. 10. Temperature dependence of conductivity for cathode V (BaO).

mined by making a Schottky plot[10] ($\log J$ *vs* $V_A^{\frac{1}{2}}$) and extrapolating the straight line portion to zero anode voltage. Conductivity and thermionic emission measurements were made not only on the same coating but at exactly the same time. This ability to take simultaneous measurements of emission and conductivity is a distinct advantage of this method.

Typical Richardson plots for the two types of coatings are shown in Figs. 12 (cathode V) and 13 (cathode II). The apparent work functions of 1.79 ev and 1.42 ev, respectively, were obtained from the slopes of the curves. Emission data over a similar temperature range were taken for cathodes I, II, IV, and VII. The results for these are included in Table II. Emission data on cathode VI were taken over a wide range in temperature and anode current in order to determine whether the Richardson plot remained a straight line over this range. Figure 14 shows this plot to be a straight line over the temperature range of 368 to 1017°K and an anode current range of more than 12 orders of magnitude. An apparent work function of 1.37 ev was determined from the slope of this curve.

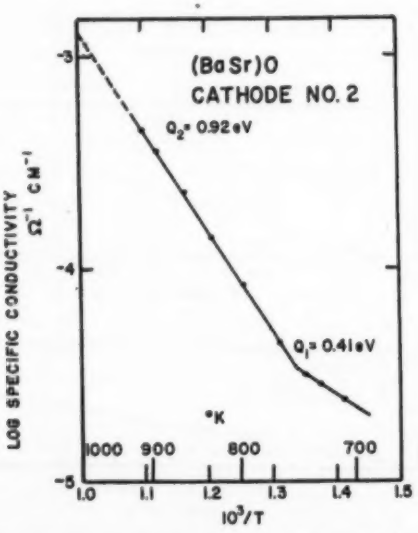


FIG. 11. Temperature dependence of conductivity for cathode II ((BaSr)O).

SUMMARY AND CONCLUSIONS

A new method of measuring the electrical conductivity has been investigated and found to give the conductivity of (BaSr)O and BaO coatings. This method has the disadvantages that it cannot be used at very low temperatures owing to the limitation set by normal current measuring devices, and can not be used for coatings whose conductivity to thermionic emission ratio is greater than 2 cm/volt. Distinct advantages are (1) the conductivity of coatings which are in a normal state for thermionic emission may be measured without the use of probe wires or other devices which might impair the thermionic emission of the sample, and (2) conductivity and thermionic measurements may be made simultaneously on the same coating sample.

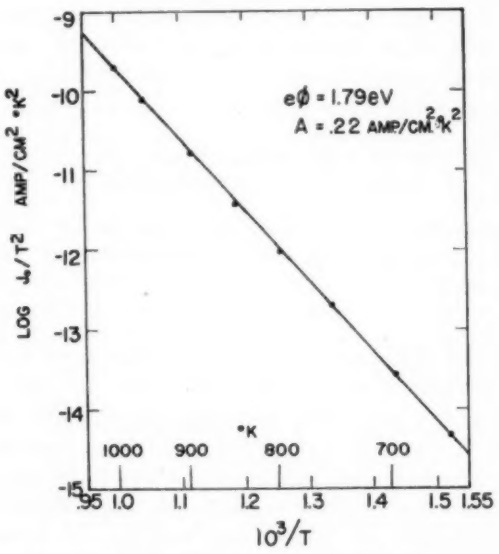


FIG. 12. Richardson plot for cathode V (BaO).

A comparison of the results on BaO with results obtained here and by others on (BaSr)O coatings indicates that the conductivity method can be used to a better advantage on the (BaSr)O coating since the latter has a lower conductivity to thermionic emission ratio. It also seems likely that this method may well find application in measuring the resistance of the interface layer formed on certain cathode base metals in which case the coating resistivity may be negligible. Some consideration has been given to the possibility

that the cathode voltages measured in this work arose from an interface resistivity rather than a coating resistivity as we assumed. This does not seem likely since (1) a pure nickel base metal was used upon which no chemical interface compounds are known to develop, (2) the conditions for interface formation should have been the same in all cathodes yet several of these showed no measurable resistivity, (3) the magnitude and temperature variation of the measured conductivity are quite similar to values reported for the oxide coating and (4) it is reported[12] that the oxide coating resistivity always exceeds that of the interface unless special base alloys are used.

On the cathodes for which the conductivity could be determined, the conductivity was measured as a function of temperature and found to yield a temperature dependence similar to that which was obtained by Loosjes and Vink and by Young. The slope of each of the two portions of the curve has been explained by Loosjes and Vink as being due to two different mechanisms in parallel. The slopes of the conductivity curve might also be interpreted as being due to two sets of energy levels with different thermal activation energies. However, this model does not seem to be plausible on a quantitative basis since the Fermi level would be required to shift abruptly at the temperature of the break. In every case the Richardson plot was found to yield only one straight line over this temperature range. In one case, a straight line was obtained over 12 orders of magnitude in anode current.

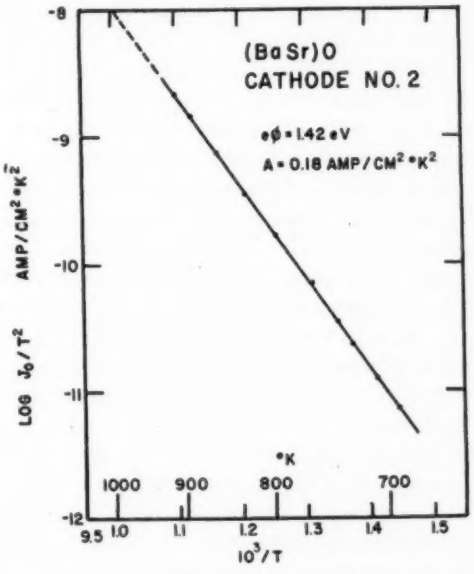


FIG. 13. Richardson plot for cathode II ((BaSr)O).

[12] C. Biguenet, Le Vide 37, 1123 (1952).

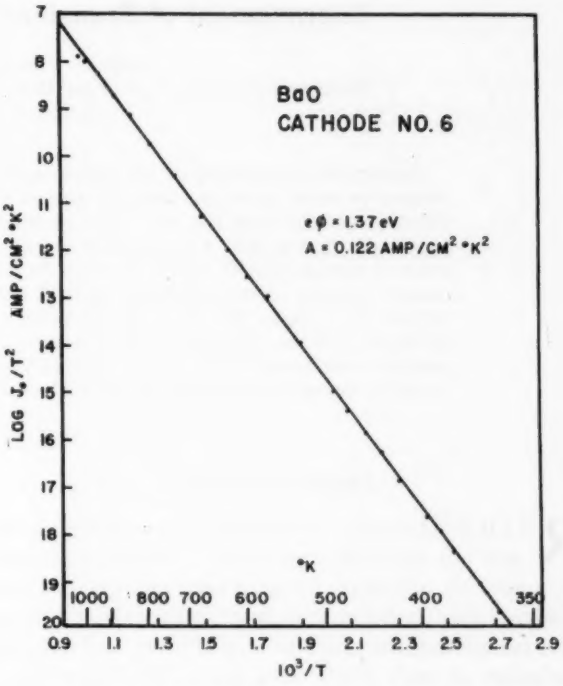


FIG. 14. Richardson plot for cathode VI (BaO).

In conclusion, the authors wish to express their appreciation to Professor A. S. Eisenstein who suggested the problem and under whose guidance the investigation was completed, and to Professor G. H. Vineyard for helpful suggestions. To the U. S. Office of Naval Research the authors are deeply indebted for the assistance which helped to make this work possible.

APPENDIX

In all of the tubes used in this investigation the ratio of anode current to saturation collector current was about 800, thus it was necessary to measure collector currents at least five orders of magnitude below the anode current to obtain the retarding potential curves. Owing to the limitation set by normal current measuring devices this factor prevented the use of this method at very low temperatures.

In an effort to increase the range of the retarding potential method two conductivity tubes were built, each incorporating a change in anode design. In the first tube the area of the apertures in the anode was increased by a factor of ten while in the second tube the anode aperture was made larger than the cathode area and it was covered with a tantalum coarse mesh screen. Only sample measurements were made on these tubes but the results indicate that the new anode designs are successful in extending the range of the retarding potential method. Plots of collector voltage as a function of anode current were straight lines as this method requires. In the first tube the ratio of anode current to saturation collector current was about 80 while in the second tube this ratio was further reduced to 1.5. Thus measurements of conductivity in the second tube could be made using collector currents only two or three orders of magnitude below the anode current. In future applications of this method it would be well to incorporate one of the above anode designs and thus increase the range and usefulness of this technique.

Engravement of Transient Stress Wave Particle Velocities*

JOHN S. RINEHART AND JOHN PEARSON
Michelson Laboratory, U. S. Naval Ordnance Test Station, Inyokern, China Lake, California
(Received November 24, 1952)

Quantitative information on the high intensity transient waves set up by impulsive loading is usually required in order to design rationally against failure under such loads. To obtain reliable data is often difficult when the loads last only a few microseconds and reach magnitudes of several hundred thousand pounds per square inch. A technique is described which enables quantitative determinations to be made either of particle velocity within the transient wave or of velocity of propagation of the wave. The scheme consists basically of measuring the depth of the permanent impression or engravement that is left on a surface when a pellet that has been previously affixed to that surface flies off. The engravement is a direct consequence of the impingement of a transient wave against the surface. Considerable numerical data are presented which establish the validity of the scheme. The simplicity of the technique suggests that it may find extensive application in impulsive loading investigations.

INTRODUCTION

RATIONAL design of equipment and materials that will not fail under impulsive or shock loads such as might be produced by explosive charges has been hindered greatly by lack of good quantitative information concerning the magnitude, duration, and nature of variation of such loads with time. The obtaining of quantitative data is usually quite difficult since, in most cases, the duration of the event is only a few microseconds and the magnitudes of the forces involved may be several hundred thousand pounds per square inch. Usually the physical quantities of most interest are particle velocity and rate of propagation of the transient disturbance induced in the body or, in other words, the spatial and temporal distributions of particle velocity both on the surfaces and within the body that has been impulsively loaded.

Optical, electronic, and mechanical methods have all been successfully used in obtaining information in specific cases. In most cases, novel approaches are required. Moreover, there appears to be no universally applicable technique. Hopkinson[1] was one of the first to study quantitatively the disturbances generated in metal rods by explosive charges and impacting missiles. His method was to measure the respective momentums of pieces which were thrown from the end of a cylindrical bar that had been struck a sharp blow at the opposite end and from these momentums to arrive at a spatial distribution of particle velocity within the disturbance. Landon and Quinney,[2] Davies,[3] and Kolsky[4] have continued with the Hopkinson bar type of investigation. A modified Hopkinson type of experiment has been used by one of the authors[5,6] in order to obtain quantitative data relating to the scabbing of metal plates under explosive attack. Generally, the Hopkinson type of experiment is most useful when the disturbances are an inch or less in length, *viz.*, for disturbances in steel of less than about five microseconds.

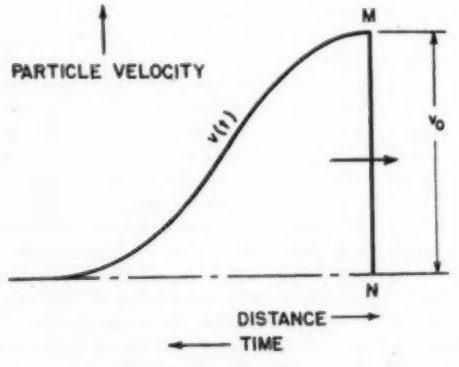


FIG. 1. General profile of transient disturbance generated by an impulsive load.

Owen and Davies,[7] Schardin,[8] Christie and Kolsky,[9] and Allen and McCrary[10] have used high speed photographic schemes for observing transient stress and particle velocity distributions.

Baldwin type strain gauges and piezoelectric crystals are sometimes used. Probes used in conjunction with high writing speed oscilloscopes have been used by some workers[11] for the measurement of velocities of propagation. Photographic and mechanical schemes are usually best suited to obtain the data that are wanted. Delay problems encountered in electronic schemes while not unsurmountable often introduce serious complications.

The present paper describes a technique which appears to show much promise in enabling quantitative

* Partially supported by the U. S. Office of Naval Research.

[1] B. Hopkinson, Trans. Roy. Soc. (London) **213A**, 437 (1914).

[2] J. W. Landon and H. Quinney, Proc. Roy. Soc. (London) **103A**, 622 (1923).

[3] R. M. Davies, Trans. Roy. Soc. (London) **240A**, 375 (1948).

[4] H. Kolsky, Proc. Phys. Soc. (London) **62B**, 676 (1949).

[5] J. S. Rinehart, J. Appl. Phys. **22**, 555 (1951).

[6] J. S. Rinehart, J. Appl. Phys. (to be published).

[7] J. D. Owen and R. M. Davies, Nature **164**, 752 (1949).

[8] H. Schardin, Glasstech. Ber. **23**, 1, 67, 325 (1950).

[9] D. G. Christie and H. Kolsky, Trans. Soc. Glass Technol. **36**, 65 (1952).

[10] W. A. Allen and C. L. McCrary, "Transient waves through steel produced by impulsive loading," paper presented at meeting of Am. Phys. Soc., Berkeley, California, December, 1951. Phys. Rev. **85**, 769(A) (1952).

[11] Pack, Evans, and James, Proc. Phys. Soc. (London) **60**, 1 (1948).

determinations to be made either of particle velocity within a transient high intensity stress wave or of the velocity of propagation of the wave. The scheme is exceedingly simple and, hence, may find application in situations where it has hitherto been impractical to make quantitative measures.

PHYSICAL BASES OF TECHNIQUE

In many cases the transient disturbance generated by impulsive loading is to a first approximation of the form shown in Fig. 1. The specific combination of variables used to specify the disturbance can be chosen in a number of different ways. The curve of Fig. 1 may be thought of as representing any one of the following: (1) a spatial distribution of particle velocity; (2) a spatial distribution of stress; (3) a temporal distribution of particle velocity; or, (4) a temporal distribution of stress. If the wave is elastic, the stress and particle velocity within the disturbance will be related through the equation:

$$\sigma = \rho c v \tag{1}$$

where σ is the stress, ρ is the density of the material, c is the velocity of propagation of the disturbance, and v is the particle velocity. From the above equation, it is evident that a measurement of particle velocity is equivalent to a measurement of stress provided the velocity of propagation of the disturbance is known and the disturbance can be considered essentially elastic. Often this latter condition is fulfilled since the disturbance is primarily an unloading wave. Only elastic disturbances will be considered here.

The particle velocity in a longitudinal wave is normal to the wave front. If the wave is one of compression, the particle velocity will be in the same direction as the direction of propagation of the wave; if the wave is one of tension, the particle velocity will be in a direction opposite to the direction of propagation of the wave.

The scheme that will be described here consists basically of measuring the depth of the engravement that is left on a surface when a pellet that has been previously affixed to that surface flies off. The engravement is a direct consequence of the impingement of the transient disturbance against the surface. The present section will be devoted to a description of what might be expected to happen at the surface.

Consider the situation shown in Fig. 2(A). The front, MN, of a wave of the type shown in Fig. 1 is moving toward the right through a semi-infinite body. The front of the wave is assumed plane and parallel to the free surface, $ABEF$. A pellet, $B'C'D'E'$, is shown affixed to the surface. The situation at a somewhat later time is shown in Fig. 2(B). The portions of the front of the wave which strike the elements AB and EF of the surface of the plate are immediately reflected back into the plate as a tension wave. The portion that strikes the surface element, BE, enters the pellet and is soon reflected from the face, $C'D'$, of the pellet.

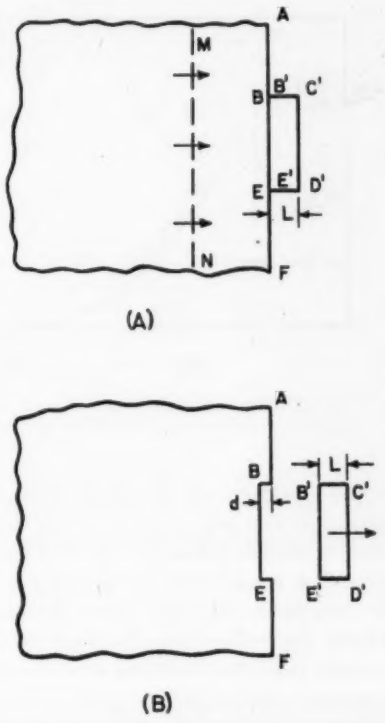


FIG. 2. Dynamics of formation of engravement as result of the impingement of a transient disturbance against a free surface.

Finally when the front of the reflected wave arrives at the interface between the pellet face, $B'E'$, and the surface element, BE, the pellet will separate from the plate and begin to move off toward the right. During the time that has elapsed between the situation in Fig. 2(A) and that in Fig. 2(B), the pellet will have left a permanent impression in the surface of the plate.

The depth of the permanent impression or engravement will depend upon the thickness of the pellet, the velocity of propagation of the wave, and the distribution of particle velocity within the wave. The velocities acquired by the several surface elements involved are shown schematically as a function of time in Fig. 3. In looking at Fig. 3 it is to be remembered that at a free surface, boundary conditions require that the particle velocity at the surface be just twice that of the incident wave at any point on the wave. Therefore, during the interval from t equal to 0 to t equal to $2L/c$, AB and EF will have twice the velocity of BE and $B'E'$. When t is equal to $2L/c$, the pellet will leave the surface and thereafter all surface elements will move with the same velocity. It is during the interval from t equal to 0 to t equal to $2L/c$ that the engravement occurs. The distance which each surface element has moved as a function of time is shown schematically in Fig. 4. If edge effects are neglected, the depth of the engravement, d, will be given by the equation

$$d = \int_0^{2L/c} v(t)dt \tag{2}$$

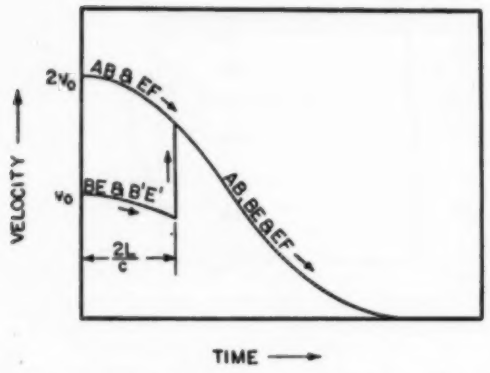


FIG. 3. Velocities of the surface elements shown in Fig. 2 during and immediately after engravement.

where L is the thickness of the pellet, c is the velocity of propagation of the wave, and $v(t)$ is the temporal distribution of particle velocity within the wave. It is evident from this equation that the depth of the engravement measures directly the average particle velocity within the wave from its front to a point $2L$ behind the front provided that the velocity of propagation of the wave is known. There must, of course, be an elastically recoverable portion which is not included in d; however, it is believed to be negligible and so dependent on edge effects as to be indeterminable. Theoretically, all that would be necessary in order to construct a particle velocity distribution would be to run a series of trials in which pellets of several different thicknesses were used and to measure the depths of the engravements made by the respective pellets. The procedure used to construct the detailed particle velocity distribution would be quite similar to that used in previous studies in which pellet velocities rather than engravement depths were measured.[5]

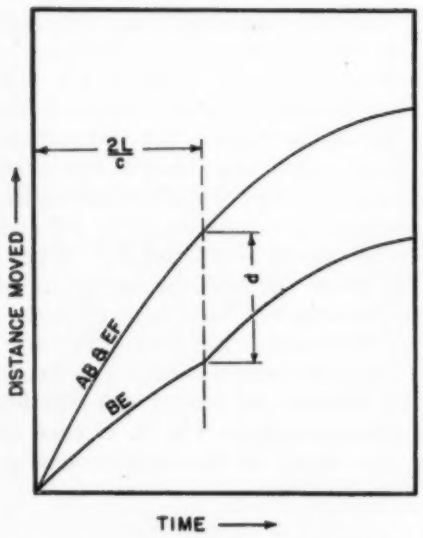


FIG. 4. Distance moved by the surface elements shown in Fig. 2 during and immediately after engravement.

It is apparent that, if both the velocity with which the pellet is thrown from the plate and the depth of the engravement are measured, then the velocity of propagation of the wave can be computed. The velocity of the pellet will be given by the equation

$$v_p = 2\int_0^{2L/c} v(t)dt \Big/ 2L/c, \tag{3}$$

and hence the velocity of the propagation of the wave will be given by the equation

$$c = (L/d)v_p. \tag{4}$$

A differential velocity will exist between the respective surface elements only for a period of time which corresponds to the wavelength of the disturbance. If the length of the transient disturbance is less than twice the thickness of the pellet, the time during which deformation occurs will be less than $2L/c$ and the depth of engravement will remain constant even though the pellet thickness is increased. The situation is illustrated schematically in Fig. 5.

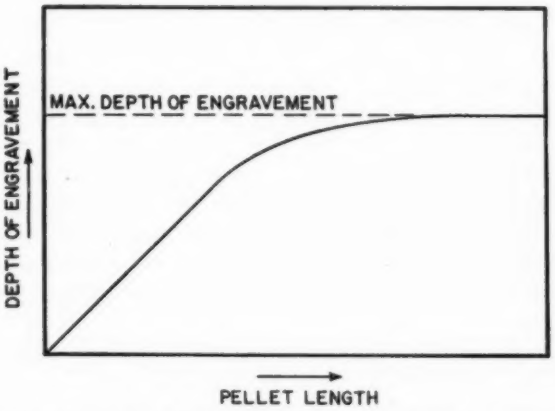


FIG. 5. Depth of engravement as length of pellet is increased.

Several hundred engravements that originated as indicated above were observed in connection with extensive investigations of scabbing of metal plates under explosive attack. The experimental set-up used in many of these studies is indicated in Fig. 6. A photograph of a typical engravement is reproduced in Fig. 7.

QUALITATIVE ASPECTS OF OBSERVED ENGRAVEMENTS

The ideas expressed in the previous section are somewhat novel and it was felt important to verify, in so far as possible, the fact that these general notions would, in fact, apply to real situations. Fortunately, during earlier scabbing investigations several hundred trials were made with the explosive-plate-pellet arrangement shown in Fig. 6. These trials have provided data sufficient to demonstrate conclusively that the present technique can be used for quantitative determination of particle

velocity in high intensity transient stress waves. Qualitative aspects of the observed engravements are discussed in the present section.

In the scabbing tests each explosive charge was 1 in. in diameter and 2 in. long. The metal plates ranged from $1\frac{1}{2}$ in. to 3 in. in thickness and were in most cases about 6 in. in diameter. Pellets were $\frac{1}{2}$ in. in diameter. Series of trials were carried out with series of pellets of respective thicknesses, $\frac{1}{16}$, $\frac{1}{8}$, $\frac{1}{4}$, $\frac{3}{8}$, and $\frac{3}{4}$ in. Metals of six different characteristics were used in the trials: 1020 annealed steel, 4130 annealed steel, annealed copper, annealed alpha-brass, 24*S*-*T*4 aluminum alloy, and 24*S*-*O* aluminum alloy.

A cursory examination of the plates indicated that the engravements were as to be expected. If the plate thickness increased, the depth of the engravement decreased for a given pellet thickness. A decrease is to be expected since the intensity of the wave becomes less as the wave moves through the plate. The effect is illustrated by the two photographs reproduced in Fig. 8.

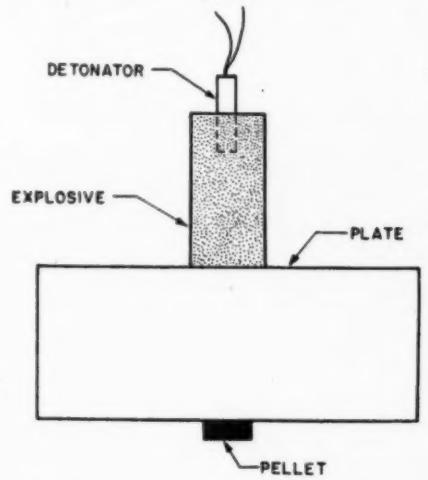


FIG. 6. Pellet-plate-charge arrangement used for scabbing and engravement studies.

For a given thickness of plate, the depth of the engravement increases with pellet thickness. The deepest engravement in any of the plates was about 0.004 in. The depths decreased from this value to no engravement at all.

Several methods were tried in an effort to obtain an accurate and detailed profile of the engravement. Use of dental plaster to make a cast of the engravement proved to be the most satisfactory. Each casting was sectioned and an enlarged (100 to 1) shadowgraph made of its cross section. The profiles of approximately 70 engravements were studied in this fashion. An example of one of the shadowgraphs is reproduced in Fig. 9. Generally, it was found that the plaster cast method made it possible to determine depths of engravements to an accuracy of about 0.0001 in. For the most part, only engravements whose depths were greater than 0.001 in. were studied in detail.

FIG. 7. Typical engravement produced in a copper plate by a copper pellet affixed to the surface of the plate, (plate: $5\frac{1}{4}$ in. diam $\times 2$ in.; pellet: $\frac{1}{2}$ in. diam $\times \frac{3}{8}$ in.; charge: 1 in. diam $\times 2$ in.).

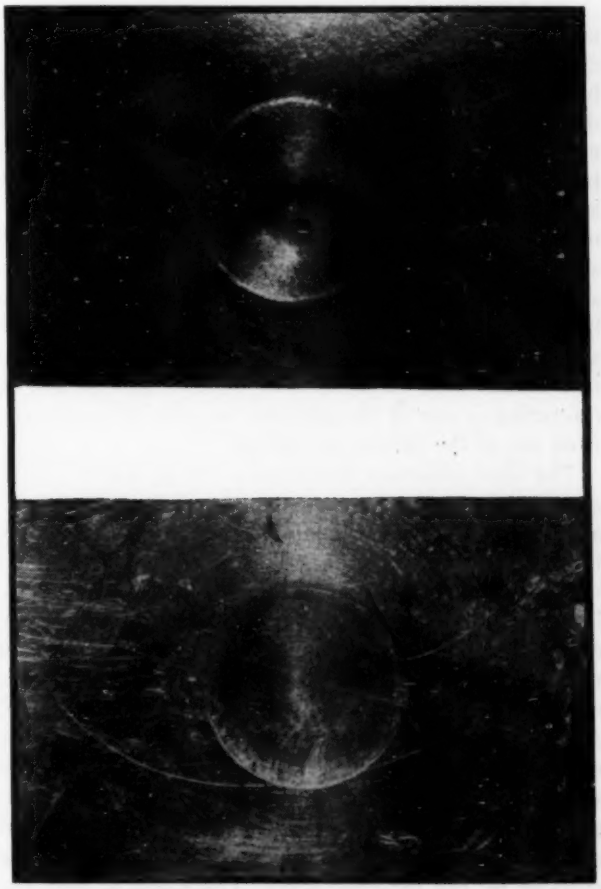


FIG. 8. Close-up views of copper plate engravements. Upper photograph: plate, 2 in. thick. Lower photograph: plate, 3 in. thick. (Pellet in each case: $\frac{1}{2}$ in. diam $\times \frac{3}{8}$ in.; charge: 1 in. diam $\times 2$ in.).

FIG. 9. Shadowgraph of casting of engravement in 24*S*–*O* aluminum alloy plate. (plate: 5½ in. diam ×2½ in.; pellet: ½ in. diam ×¾ in.; charge: 1 in. diam ×2 in.).

An interesting and extremely significant aspect of an engravement made by a pellet is illustrated schematically in Fig. 10. Here it is assumed that the plate scabs; however, the engravement was completed before the deformation of the free surface took place so that the engravement is preserved even though the surface has been severely deformed. It has been possible in many such cases to make accurate particle velocity determinations. An example of this is given in Fig. 11 which shows portions of the engravement left on the petaled surface of a copper plate. Even though complete scabbing has occurred, the depth of the engravement is still measurable. The ability to obtain information from test bodies that have suffered severe deformation or fracturing is one of the greatest benefits to be derived from use of the present technique.

The ideal situation described in the previous section cannot, of course, exist near the edge of the pellet since, as the wave moves through the pellet, the pellet will begin to expand laterally. This lateral expansion introduces certain complications but does not seem to interfere seriously with a fairly accurate determination of engravement depth. The situation near the edge of the pellet is illustrated schematically in Fig. 12. Particle motion will be substantially that shown in the figure. As the longitudinal wave moves through the pellet, the edges of the pellet begin to expand laterally. At the same time, the material in the plate that lies immediately underneath the pellet will remain in a state of compression while that outside of the pellet will be relieved by the reflected wave. The result is that the material underneath the pellet will begin to expand into the region that is relieved by the reflected wave. The over-all result will be that the edge of the engravement will slope and that its diameter will be greater than the original diameter of the pellet. An examination of engravements showed that both effects were observable.

The exact shape of the edge must depend in a complicated way on several factors. Among these factors will be Poisson's ratio, and plastic properties of the materials. The edge effects have not yet been studied in sufficient detail to warrant extensive discussion of them here. It is interesting, however, to point out that the shape of the edge of the engravement is not the same for different metals. Edge shapes typical of copper, 4130 steel, brass, and 24*S*-*O* aluminum alloy are illustrated in Fig. 13. Many of the copper engravements have a lip which indicates that a greater amount of plastic flow has taken place than for the other metals. The 4130 steel engravements have a gently sloping edge and seem to show little, if any, lip. As the thickness of the pellet is increased, there is more time for lateral expansion to occur before the pellet leaves the plate, and an increase in amount of lateral expansion with pellet thickness was observed in a series of 4130 steel engravements. The cleanest and best engravements were made in 24*S*-*O* aluminum alloy.

QUANTITATIVE OBSERVATIONS

Quantitative verification of the technique has consisted specifically of using previously measured pellet velocities and the depths of the engravements to com-

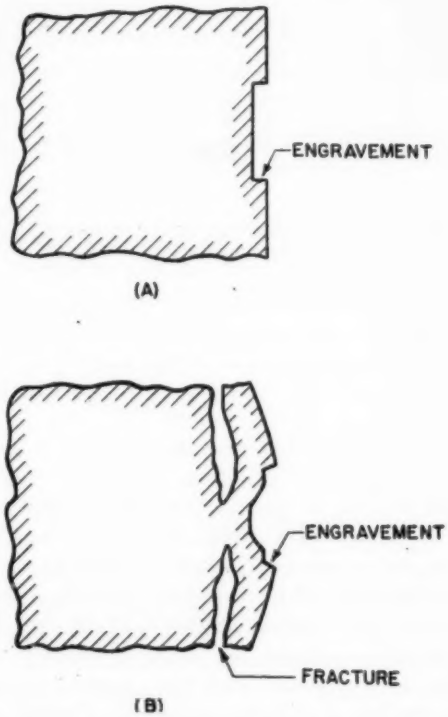


FIG. 10. Schematic cross-sectional views showing engravement in scabbed and deformed plate. (A) Engravement complete, plate not yet scabbed. (B) Plate scabbed and deformed, engravement preserved.

pute the velocity of the propagation of the disturbance from Eq. (4). Individual values of c so obtained are plotted in Fig. 14 for several different metals. In measuring the depth of the engravement, the distance d of Fig. 13 was used. Two measurements have been made on most engravements, one near each edge of the cross section of the casting. The two readings were averaged in plotting each point in Fig. 14. The engravements on the 1020 steel, 4130 steel, and 24S-T aluminum alloy plates were so poor that accurate measurements could not be made on them. Reference values of c are shown as dashed lines in the figure. The reference

FIG. 11. Scabbed copper plate showing preservation of pellet engravement. (Plate: 5½ in. diam ×1½ in.; pellet: ½ in. diam ×½ in.; charge: 1 in. diam ×2 in.).

values are those which were obtained from previous studies[12] by the authors in which conical surfaces of fracture were used to determine wave velocities. The agreement between the present values and the reference values is sufficiently good to establish confidence in the validity of the present technique.

There is considerable spread in the values plotted in Fig. 14. It is interesting to examine the probable origins of the spread. First, it should be remembered that the data plotted here are obtained from plates which had originally been used for another purpose and, hence, no particular precautions were taken to insure excellence of engravement. For example, a rubber base

[12] J. S. Rinehart and J. Pearson, J. Appl. Phys. 23, 685 (1952).

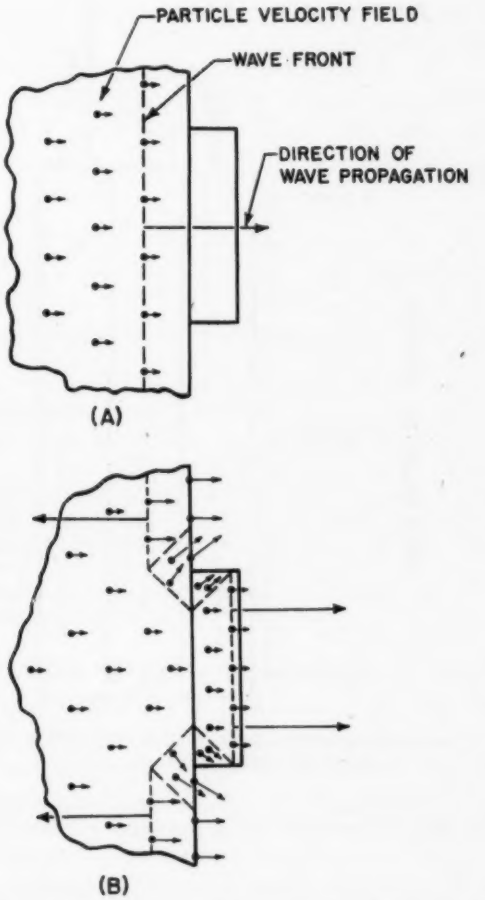


FIG. 12. Sketch indicating certain qualitative features of particle velocity fields that contribute to edge effects.

cement was used to affix the pellet to the plate. The thickness of this layer of cement might perhaps be 0.0005 in. This could easily introduce an error of 10 to 20 percent in measurement of velocity of propagation. Small quantities of foreign matter such as dust and sand could, in some cases, have been lodged between the pellet and the plate. Roughness of the plate

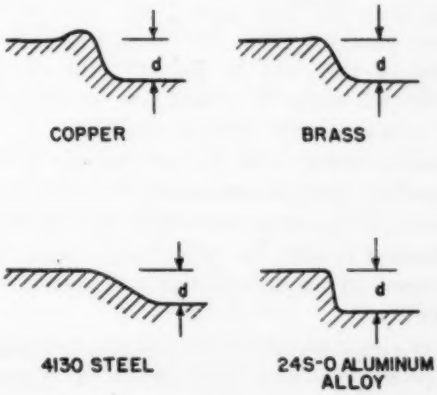


FIG. 13. Cross-sectional views showing typical edge contours for several metals.

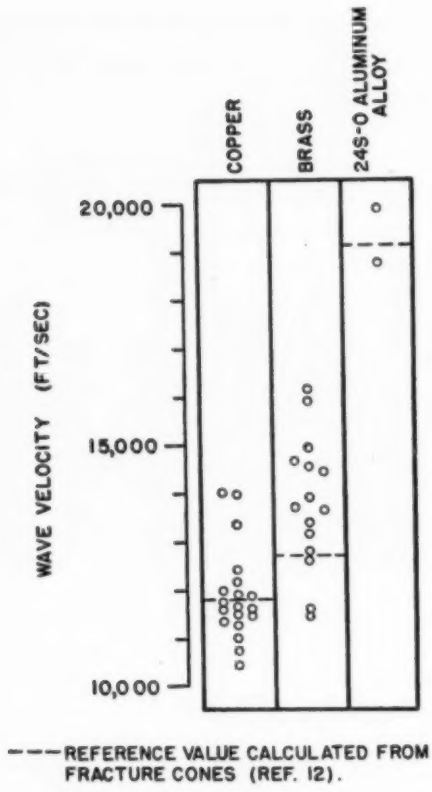


FIG. 14. Computed values of wave velocities obtained from measurement of depth of engravement.

surface could also contribute to poor engravement. A burr along the edge of a pellet or a breaking of the edge of a pellet during machining could also introduce certain errors. Second, the plaster castings made of the engravements were sanded with fairly coarse No. 400 sandpaper so that the roughness of the casting in some cases amounted to about 0.0004 in. Casting roughness is apparent in the shadowgraph of Fig. 9. In a few cases, the plates had rusted or had accumulated dirt so that a truly accurate impression of the engravement could not be obtained. All of the above factors would contribute toward obtaining depths of engravements which were less than the true depth, that is, would lead to computed wave velocities that were too high. An examination of the data in Fig. 14 indicates that in practically all cases in which the computed wave velocity is in error, the value is too high.

The above errors could be considerably reduced by taking extreme care in machining the contact surfaces and pellets and by using something more suitable than rubber cement to affix the pellet to the plate. Use of a thin grease or magnetization of pellet suggest themselves.

Another check of the validity of the technique is to predict pellet velocities from depth of engravement

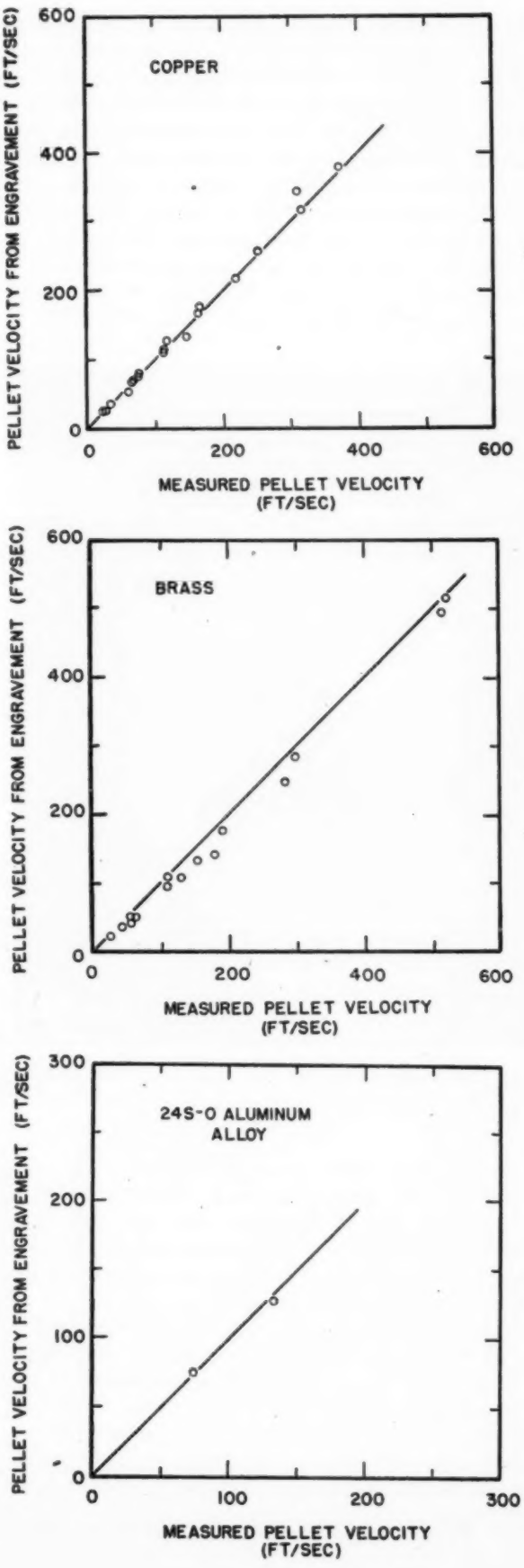


FIG. 15. Graphs showing correlation between pellet velocities deduced from engravements and measured pellet velocities. Curve in each case is straight line having slope of one.

and an assumed value of wave velocity and then to compare this predicted value of pellet velocity with pellet velocities which were previously measured. Such a comparison is shown for copper, brass, and 24S-*O* aluminum alloy in Fig. 15. The straight line drawn in each of the three plots has a slope of one and passes through the origin. The cluster of the points about the line indicates that there is good agreement between predicted and measured pellet velocities.

In conclusion, the authors wish to acknowledge the assistance of Mrs. Audrey Giger who made many of the plaster castings.

JOURNAL OF APPLIED PHYSICS VOLUME 24, NUMBER 4 APRIL, 1953

Photoelastic Studies of Quenched Cylinders and Spheres

K. A. Parsons
Department of Physics, Michigan State Normal College, Ypsilanti, Michigan
(Received December 22, 1952)

A mathematical theory of residual stresses of quenching origin based on an idealized model of the quenching process is investigated. Photoelastic curves are computed from the theory and compared with those determined with a Babinet Compensator. Ten glass cylinders and spheres quenched in molten lead are employed. Although the qualitative agreement between theory and experiment is good, a quantitative comparison shows that the theory needs improvement. Major factors in the disagreement are probably the neglect of the effects of plastic flow and the temperature dependence of the physical constants.

INTRODUCTION

A SERIES of quenching experiments on glass cylinders and spheres has been undertaken in order to evaluate a recent mathematical theory of quenching stresses based on an idealized model of the quenching process.[1] This theory is essentially a reinterpretation of the classical theory of thermal stresses. The effects of plastic flow and the temperature dependence of the physical constants are neglected. Equations are derived which predict the characteristic photoelastic patterns to be expected from quenched cylinders and spheres. It has been shown previously that the general shapes of the theoretical patterns for solid cylinders agrees qualitatively with the corresponding experimental patterns.[2] In the present investigation, theoretical and experimental curves are compared in a quantitative manner. The theoretical photoelastic curves are computed, employing the observed quenching temperatures together with the physical constants of the glass. The corresponding experimental curves are determined with a Babinet Compensator. This is done for seven solid glass cylinders and three glass spheres in a variety of sizes quenched in molten lead. A wide range of quenching temperatures is employed.

The experimental photoelastic curves presented here, may be employed with the above mathematical theory of quenching stresses to determine the distribution of the separate principal stresses in glass cylinders and spheres. This aspect of the problem is not considered here, but will be presented in a future paper.

[1] R. C. O'Rourke and A. W. Saenz, Quart. Appl. Math. 8, 303 (1950).

[2] Everett, Parsons, Saenz, and O'Rourke, *Monthly Progress Reports*, Engineering Research Institute, University of Michigan, Ann Arbor, 1946–1948. Unpublished.

RESUME OF THE PHOTOELASTIC EQUATIONS

The photoelastic equations for quenched cylinders and spheres are derived by employing the stress-optic law and the theory of thermal stresses together with an idealized model of the quenching process. The stress-optic law is employed to find the retardation produced by the principal stresses in the plane of the wave front. The expression for the integrated retardation of a plane monochromatic wave which traverses a cylinder or sphere is

$$R(\eta)=C\int_{-\eta}^{+\eta}(P-Q)dy, \tag{1}$$

where $R(\eta)$ is the total retardation, C is the photoelastic coefficient, dy is the length of the light path, and P and Q are the principal stresses in the plane of the wavefront. It is shown from the theory of elasticity that for a quenched long cylinder whose surface is maintained at a temperature T_0, $P=\sigma_z$ and $Q=\cos^2\phi\sigma_\rho+\sin^2\phi\sigma_\phi$, where σ_z is the axial stress, σ_ρ is the radial stress, and σ_ϕ is the circumferential stress. Thus,

$$R(\eta)=C\int_{-\eta}^{+\eta}\{\sigma_z-\cos^2\phi\sigma_\rho-\sin^2\phi\sigma_\phi\}dy. \tag{2}$$

Inasmuch as σ_z, σ_ρ and σ_ϕ are functions of ρ only, Eq. (2) is simplified by expressing $\cos^2\phi$, $\sin^2\phi$, and y in terms of ξ and ρ. Then,

$$R(\xi)=2C\int_\xi^b\left\{(\sigma_z-\sigma_\phi)+(\sigma_\phi-\sigma_\rho)\frac{\xi^2}{\rho^2}\right\}\frac{\rho d\rho}{(\rho^2-\xi^2)^{\frac{1}{2}}}. \tag{3}$$

The corresponding equation for a sphere is

$$R(\xi)=2C\xi^2\int_\xi^b(\sigma_\phi-\sigma_r)\frac{dr}{r(r^2-\xi^2)^{\frac{1}{2}}}. \tag{4}$$

TABLE I. Physical constants of lime glass.

C	Photoelastic coefficient	2.80×10^{-12} cm²/dyne
α	Linear coefficient of expansion	9.78×10^{-6}/°C
E	Young's modulus	6.48×10^{11} dynes/cm²
ν	Poisson's ratio	0.18
χ	Diffusivity	4.67×10^{-3} cm²/sec
T^*	Annealing temperature	533°C

A simplified model of the quenching process is postulated in order to permit the calculation of σ_ρ, σ_ϕ, and σ_z from the theory of thermal stresses. It is assumed that residual stresses of quenching origin are essentially elastic, and that when a piece of glass which is heated uniformly to its softening temperature, T_1, is suddenly cooled in an infinite bath at T_0, elastic stresses begin to form and remain permanently set in the glass as soon as the critical temperature T^* is reached. ($T_1 > T^* > T_0$.) The effects of plastic flow, and the temperature dependence of the physical constants of the glass are neglected. These assumptions lead to the following rule:

To calculate residual stresses in a body (A) quenched from the uniform temperature T_1 to the temperature T_0 of a bath ($T_1 > T_0$), calculate the thermal stresses at time

$$t = t^*(T = T^*) \tag{5}$$

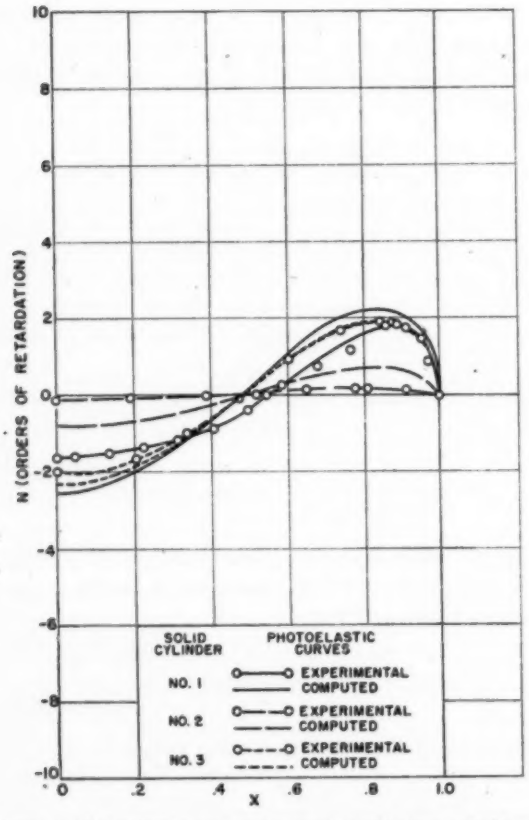


FIG. 1. Photoelastic curves for small solid glass cylinders.

in an elastic body (B), with the same elastic and thermal constants as (A), whose initial temperature is T_1 and whose surface, beginning at $t=0$, is maintained at T_0. The stresses in (B) at $t=t^*$ are the negative of the residual stresses at $t=\infty$ in (A).

The solutions for σ_ρ, σ_ϕ and σ_z are taken from the theory of thermal stresses with signs changed and t^* replacing t according to Eq. (5), and are substituted in Eqs. (3) and (4) to obtain the theoretical photoelastic

TABLE II. Experimental data.

	Radius ϕ (cm)	Quenching temperature T_0 (°C)
Solid cylinders		
No. 1	0.631	377
No. 2	0.632	482
No. 3	0.632	396
No. 4	2.85	488
No. 5	2.60	421
No. 6	2.50	449
No. 7	2.50	417
Spheres		
No. *A*	5.11	450
No. *B*	4.65	523
No. *C*	4.80	491
λ	Wavelength of sodium doublet	5.893×10^{-5} cm
T_1	Furnace temperature	627°C

equations. Equation (3) for solid cylinders becomes

$$R(x) = -\frac{4C\alpha E(T_1-T_0)}{1-\nu}\cdot b\cdot\sum_{n=1}^{\infty}\frac{\exp^{-\chi\delta_n^2 t^*}}{\delta_n^2 b^2 J_1(\delta_n b)}\cdot R(n;x), \tag{6}$$

TABLE III. The effect on the theoretical photoelastic curves of increasing the value of T^*.

Solid cylinder	Times increase in retardation for $T^*=550$°C	Times increase in retardation for $T^*=575$°C
1	1.11	1.30
2	1.33	1.85
3	1.14	1.33
4	1.39	1.96
5	1.16	1.40
6	1.21	1.52
7	1.15	1.39
Spheres		
A	1.22	1.57
B	2.71	5.29
C	1.41	2.04

where

$$R(n;x) = -J_1(\delta_n b)(1-x^2)^{\frac{1}{2}} + \sum_{\kappa=0}^{\infty}\sum_{l=0}^{\infty}\frac{(-)^{\kappa+l}(2l+1)(\delta_n b)^{2\kappa+2l}}{4^{\kappa+l}\cdot\kappa!(\kappa+2l+1)!}\cdot(1-x^2)^{\kappa+l-\frac{1}{2}}\cdot\frac{J_{2l+1}(x\delta_n b)}{x}\cdot\left\{\frac{1-x^2}{2(\kappa+l)+1}+\frac{2(\kappa+l)x^2}{2(\kappa+l)-1}\right\};$$

δ_n are positive roots of $J_0(\delta b)=0$; $x=\xi/b$.

Equation (4) for spheres becomes

$$R(x)=-\frac{4C\alpha E}{1-\nu}(T_1-T_0)\left(\frac{\pi^3}{2}\right)^{\frac{1}{2}}\cdot b \cdot\sum_{n=1}^{\infty}(-)^n\cdot n\cdot\exp\left(-\frac{n^2\pi^2\chi t^*}{b^2}\right)R(n;x), \quad (7)$$

where

$$R(n;x)=x^2\sum_{\kappa=0}^{\infty}\frac{\pi^{\kappa}(2n)^{\kappa-\frac{1}{2}}\kappa!}{(2\kappa+1)!}(1-x^2)^{\kappa+\frac{1}{2}}J_{\kappa+5/2}(n\pi).$$

For convenience in comparing the theoretical photoelastic curves with data as taken with a Babinet Compensator, Eqs. (6) and (7) are converted to orders of retardation, $N(x)$, where $N(x)=R(x)/\lambda$, and λ = wavelength of the sodium doublet.

The values of the physical constants of the glass are listed in Table I. The dimensions of the glass specimens and the quenching temperatures are listed in Table II.

REMARKS ON THE PHOTOELASTIC CURVES

The computed photoelastic curves are compared with the corresponding experimental curves in Figs. 1–3. The agreement is good for cylinder No. 3 (Fig. 1) and for sphere B (Fig. 3). The curves are close for cylinder No. 5 (Fig. 2). But a general disagreement between theory and experiment is apparent among the other curves, being especially large for cylinder No. 2 (Fig. 1), cylinders Nos. 4, 6, and 7 (Fig. 2), and for spheres A and C (Fig. 3). The cylinders in Fig. 1 are about 1.25 cm in diameter, those in Fig. 2 about 5 cm in diameter, and the spheres in Fig. 3 about 10 cm in diameter.

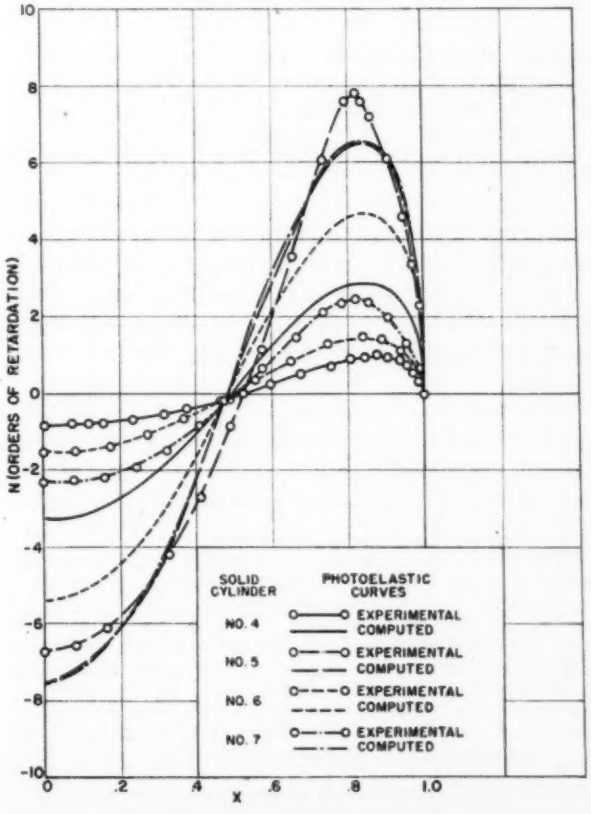


FIG. 2. Photoelastic curves for large solid glass cylinders.

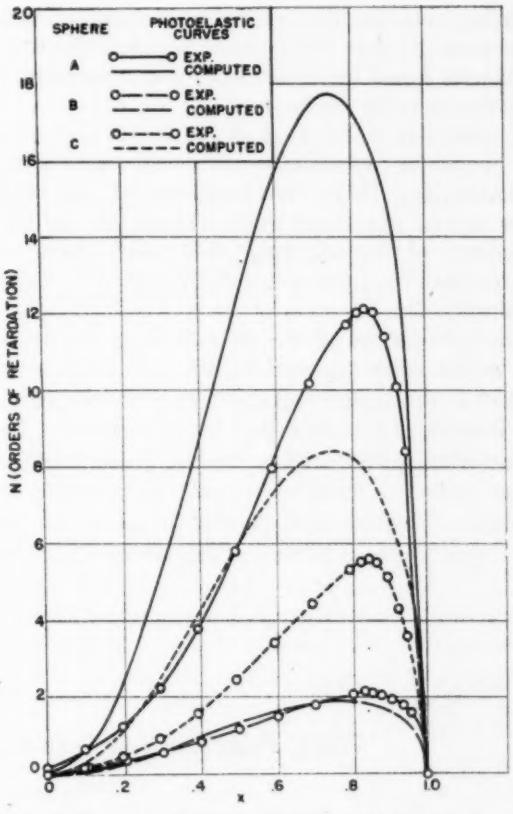


FIG. 3. Photoelastic curves for large solid glass spheres.

The disagreement is brought out numerically by comparing the ratios of the ordinates and areas of the theoretical *vs* experimental curves for all of the cylinders and spheres. The ratio of the average maxima of the computed *vs* experimental curves is 1.40 for the three spheres and 1.50 for the seven solid cylinders. The ratio of the average minima of the computed *vs* experimental curves is 1.90 for the seven cylinders. When the corresponding areas are compared, the sum of the areas under the computed curves is found to be 2.6 times the sum of the areas under the experimental curves. The ratio of the areas of the computed *vs* experimental curves is 1.60 for the three spheres.

Inasmuch as the theory requires that the areas under the maximum and minimum parts of the photoelastic curves for the solid cylinders be equal [$(\int_a^b R^*(x)dx=0$, Eq. (8)], this was employed to make an estimate of the magnitude of the experimental photoelastic errors. These areas were measured and compared for the seven solid cylinders, for both the computed and experimental curves. The maximum and minimum areas differed by only 1.5 percent for the computed curves. The corre-

sponding areas for the experimental curves differed by 10 percent. It does not appear then that the large discrepancies found between theory and experiment come from photoelastic errors.

A quantitative estimate of the errors introduced in the quenching operations cannot be made from the available data. There were fluctuations in the temperatures of the quenching baths. However, the circular symmetry of the resulting photoelastic patterns indicates that the cylinders and spheres were quenched uniformly.

The calculations of the theoretical curves are sensitive to the value assumed for the critical temperature T^* and to the experimental values of Young's modulus and Poisson's ratio. In regard to the latter two, further experimental work needs to be done to determine their values more accurately over the temperature ranges employed. There is also uncertainty as to the correct value of T^*. The annealing temperature of lime glass (533°C) was arbitrarily selected as the critical temperature. Inasmuch as quenching stresses are formed at 523°C (see Table II), it would seem that 533°C is as low a value for the critical temperature as could be justified. On the other hand, as higher values are taken for T^*, the disagreement between theory and experiment rapidly worsens. For example, at 550°C the ordinates of the theoretical curves are 2.7 times larger than at 533°C; and at 575°C they are 5.3 times larger than at 533°C, in the case of sphere No. B (see Table III).

ACKNOWLEDGMENT

This investigation was sponsored by the Department of Engineering Mechanics and the Engineering Research Institute of the University of Michigan. I am especially grateful to Professors E. L. Eriksen and F. L. Everett for the use of the Photoelastic Laboratory, and for their encouragement of this work, both official and personal.

JOURNAL OF APPLIED PHYSICS VOLUME 24, NUMBER 4 APRIL, 1953

Work Functions of Gas-Coated Tungsten and Silver Surfaces

G. L. WEISSLER AND T. N. WILSON
Physics Department, The University of Southern California, Los Angeles 7, California

(Received January 8, 1953)

The average electronic work functions of tungsten and silver surfaces, subjected to glow discharges in various gases (He, A, H_2, N_2, O_2, freon-12) and to water vapor, have been determined by Oatley's magnetron contact potential method. Oxygen and freon-12 produced increases in work functions in excess of 1 ev, whereas nitrogen and particularly hydrogen caused decreases of the order of 0.5 ev. The inert gases and water vapor produced relatively little changes. The history and treatment of the surfaces is described, and the data presented have been found to be reproducible.

INTRODUCTION

IN gas discharges certain important mechanisms are active at the electrodes, such as the liberation of secondary electrons by (a) primary electrons, (b) positive ions, (c) photons, and (d) metastable atoms. These processes have been discussed in detail by Loeb[1] and have been shown to depend on the work function ϕ of the electrode surface. Values of ϕ for many pure metals have been measured.[2,3] However, they are usually not the ones to be associated with the electrodes which are exposed to a gas discharge, because the very presence of a gas alters surface conditions materially, particularly those of cathodes which are subjected to positive ion bombardment. The role of the cathode in discharge instability[4] has been investigated with reference to changes in ϕ during the discharge, particularly with respect to onset phenomena. For reasons such as these it seemed important to study the effects of various surface treatments on the *average* electronic work function.[5]

It became clear in the early phases of this work that greater stability and reproducibility of measurement could be achieved if the surfaces were subjected to bombardment by ions of a specific gas in a glow dis-

[1] L. B. Loeb, *Fundamental Processes of Electrical Discharge in Gases* (John Wiley and Sons, Inc., New York, 1939), p. 403.

[2] For a review of work functions and methods of determination, see H. B. Michaelson, J. Appl. Phys. **21**, 536 (1950); A. L. Hughes and L. A. Du Bridge, *Photoelectric Phenomena* (McGraw-Hill Book Company, Inc., New York, 1932); S. Dushman, Revs. Modern Phys. **2**, 381 (1930).

[3] See also T. N. Wilson, Master's thesis, The University of Southern California, August 1950.

[4] L. B. Loeb, Phys. Rev. **76**, 255 (1949).

[5] Effects of surface treatment on ϕ have been measured among others by: Richardson and Cooke, Phil. Mag. **20**, 173 (1910), for Pt+O or H; C. C. Van Voorhis, Phys. Rev. **30**, 318 (1927), for Mo+A, H_2, N_2; R. Sulermann, Z. Physik **33**, 63 (1925), for Ag+H_2, O_2, etc; Hughes and Du Bridge, *Photoelectric Phenomena* (McGraw-Hill Book Company, Inc., New York, 1932); p. 78; O. Klein and E. Lange, Z. Elektrochem. **44**, 542 (1938), for summary of ϕ's of gas-coated metals; W. E. Meyerhof and P. H. Miller, Rev. Sci. Instr. **17**, 15 (1946), for W washed with H_2O, etc.; I. Langmuir and K. H. Kingdon, Phys. Rev. **34**, 129 (1929), for W+Cs, Th, O; R. C. L. Bosworth, Proc. Roy. Soc. (London) **A162**, 32 (1937), for W+alkali metals, H, O; C. W. Oatley, Proc. Roy. Soc. (London) **A155**, 218 (1936), Proc. Phys. Soc. (London) **51**, 318 (1939), for Pt+H, O, etc.

charge in contrast to simply exposing the metal to gases at various pressures. Consequently, the measurements reported here concern themselves with values of ϕ after the clean surfaces of W and Ag had been exposed to glow discharges in electropositive gases, H_2 and N_2, electronegative gases, O_2 and freon-12, inert gases, A and He, and to water vapor. Such surfaces correspond more realistically to the character of cathodes operating in discharges.

APPARATUS AND EXPERIMENTAL PROCEDURE

With slight modifications the method of measurement followed that of Oatley.[6] Briefly it depends upon the magnetron action of an axial magnetic field upon the electron current from a W filament, the reference surface, to the collector in a cylindrical diode. The anode consists of the metal or treated surface whose contact potential difference, cpd, with respect to the W filament, is to be found. The following potential relationship has been shown[6] to hold to a sufficient degree of approximation: $V=V_a+V_{cpd}+V_T$, where V is the total potential existing between anode and cathode, V_a is the applied potential, V_{cpd} is the cpd between the hot W wire and the anode surface, and V_T is a term which allows for the average kinetic energy of the electrons emitted from the filament.

If H_0 is the magnetic field which, when applied parallel to the axis of the diode, reduces the anode current to one-half its original value, then it was shown by Hull[7] that $V=H^2R^2e/8m$, where R is the radius of the anode. Since H_0, which is furnished in this experiment by Helmholtz coils, is proportional to the current I_0 producing it, $H_0=kI_0$, then one can write $V_a+V_{cpd}+V_T=k'I_0^2$, where k' is a constant characteristic of the geometry. If, for a given anode surface and filament temperature, readings of the 50 percent cutoff current, I_0, are observed for several applied potentials, V_a, then a graph of I_0^2 plotted against V_a will be a straight line with the intercept on the V_a axis equal to $-(V_{cpd}+V_T)$. Using the appropriate magnitude[6] of V_T (about 0.2 ev), the value of V_{cpd} can be found. Since V_{cpd} is of the order of one volt, the maximum V_a must not be much greater than about 10 volts in order to have sufficient accuracy in V_{cpd}. Since $\phi_{cathode}-\phi_{anode}=V_{cpd}$, the anode work function, ϕ_{anode}, can be estimated using the experimentally determined V_{cpd} and the best literature value of the work function for clean W, namely $\phi_{cathode}=4.67$ ev. This assumes that the work function of the filament was that of clean W during the measurements, which may be in error. Since, however, the changes of such work functions under treatment here are of importance, the procedure of using a reference value seems justified.

A schematic diagram of the apparatus is shown in Fig. 1. The entire assembly was placed in a Pyrex tube,

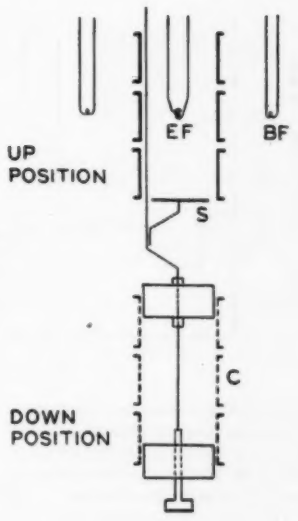


FIG. 1. Schematic of electrode assembly. *EF* and *BF*: evaporation and bombardment filaments; *s*: evaporation shield; *c*: collector electrode.

which could be dismounted easily, and joined to an oil diffusion pump. All the proper precautions[6] were taken before and during the measurements. The collector assembly, consisting of a central collector and two guard cylinders all made of Ta, could be lifted magnetically from a *Down* position, where the cpd determinations were made, to an *Up* position for surface treatment. When in this latter position the assembly was symmetrically placed with respect to three tungsten filaments. The center one could be used to evaporate a clean silver surface onto the cylinders, or it could become the anode in a low pressure glow discharge with the collector cylinder as cathode. The two outer filaments could be made the cathodes with respect to the collector so as to bombard it with 200-ma electron current at about 600 volts energy to red heat. In this way it was possible to subject the collector to various types of surface treatments and then measure their effect on ϕ when the assembly was in the *Down* position.

A mechanical commutator alternated at about 50 cycles per second the applied anode potential V_a with the heater current through the magnetron filament so that the heater current was off while measurements were taken, thereby avoiding the error caused by a voltage drop along the filament. The current I_0 was obtained by plotting a cut-off curve for each value V_a, and five to six values of V_a differing by 1.5 to 2 volts were considered sufficient for an accurate straight-line plot of a V_a *versus* I_0^2 from which V_{cpd} was determined.

The following procedure was used for a given metal surface. Metal was evaporated onto the collector, and its work function was measured at least twice and often four times. Then the surface was treated with a glow discharge in one of various gases and the work function was taken again. The collector was then restored to its initial condition either by heating it to red heat with electron bombardment or by a glow discharge in He or A, and the work function was rechecked. More frequently, however, the above processes were repeated.

[6] C. W. Oatley, Proc. Roy. Soc. (London) A155, 218 (1936), Phys. Soc. (London) 51, 318 (1939).
[7] A. W. Hull, Phys. Rev. 18, 31 (1921).

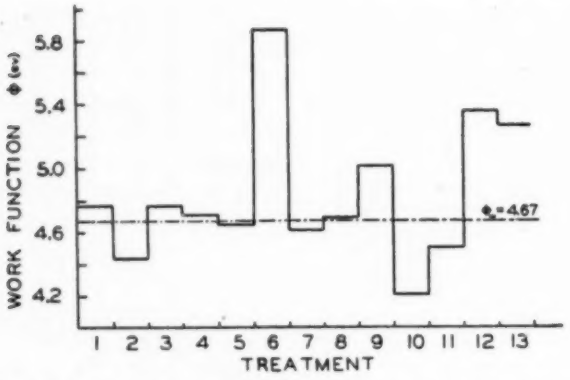


FIG. 2. Work functions of a tungsten surface. The treatments are (1) electron bombardment at 1100°C, (2) N_2 glow discharge at 700°C, (3) electron bombardment, (4) argon glow discharge at 800°C, (5) electron bombardment, (6) O_2 glow discharge, (7) argon glow discharge, (8) and (9) electron bombardment, (10) and (11) H_2 glow discharge, (12) and (13) electron bombardment. See text for detailed discussion.

The gases used for treatment were taken in increasingly electronegative order. Both heating by electron bombardment and by ion bombardment were reasonably successful, but not reliably so, in cleaning surfaces which had been exposed to the electropositive gases N_2 and H_2. Both methods failed in the cases of O_2 and freon-12 treatment.

EXPERIMENTAL DATA AND DISCUSSION

Figure 2 represents the work functions of a tungsten surface which had been obtained accidentally when the W filament of the magnetron burned out several times in a high vacuum as a result of high temperature flashing before trial runs. Each horizontal line to the left of the highest one (O_2 treatment) is characteristic of two separate work function determinations both lasting two hours each. The internal consistency was about 0.05 volt or less. The horizontal lines to the right of the maximum are the result of one ϕ measurement each. In addition, a good literature value for pure tungsten,[7] $\phi_W = 4.67$ ev, is indicated by a dot-dash line. After the tungsten evaporation, the surface treatments proceed from left to right in Fig. 2. (1) The collector was bombarded with electrons for 30 minutes total and thereby outgassed at about 1100°C. This temperature limit was determined by the adjacent Pyrex envelope. (2) A low pressure (10 mm Hg) N_2 glow discharge was operated for two minutes, heating the collector to 700°C. This lowered the work function by 0.35 ev. (3) Outgassing by electron bombardment as in (1) restored the surface to its initial condition. (4) An argon glow discharge at 35-mm Hg, which heated the surface for two minutes to 800°C, did not cause any material change in ϕ and can therefore be used for cleaning and outgassing purposes. (5) Further electron bombardment at 1100°C produced very little difference. (6) A very weak glow discharge of 10 ma at 650 volts in O_2 was used in contrast to a strong one to keep the cylinder cool and prevent the formation of visible oxides. This raised the work function by 1.20 ev. (Oatley[6] found that O_2 on platinum increased the work function of pure Pt by 1.20 ev.) (7) After the O_2 treatment, the W surface was again bombarded with a strong argon glow discharge, with the surprising result that the initial low work function (1) was restored. (8) Subsequent electron bombardment at 1100°C did not change this much. (9) After two days standing in a high vacuum with all traps immersed in liquid nitrogen, the surface was again bombarded with electrons, but the work function was about 0.35 ev higher than previously. However, this was not known at the time of measurement. Otherwise, argon bombardment would have been resorted to. (10) A subsequent weak glow discharge of 20 ma at 550 volts in H_2 (intense discharges caused flash over in undesirable portions of the tube) lowered ϕ by 0.85 ev. (Oatley[6] found that H_2 on platinum decreased the work function of pure platinum by 1.01 ev.) (11) Two days later the H_2 bombardment was repeated and ϕ was found to have increased by 0.30 ev. (12) and (13) Further outgassing by electron bombardment at 1100 C for 30 minutes each yielded an additional increase of about 0.8 to 0.9 ev and was therefore not successful in restoring the surface to its initial condition (1). The erratic work function changes after the O_2 treatment (6) may perhaps be explained by a continuous gradual diffusion of O to the surface.

Figure 3 summarizes the results of 74 work function determinations obtained under widely varying conditions on five separately distilled silver surfaces. Without treatment these surfaces yielded values of ϕ as measured immediately after evaporation and after standing one or two days in a high vacuum, which ranged from 3.8 to 4.5 ev with an algebraic average of 4.21 ev. For comparison purposes the most recent mean value of silver[8] is $\phi_{Ag} = 4.28$ ev. The aging in vacuum of about 1×10^{-6} mm Hg, in almost all cases, led to a decrease of the work function of about 0.1 ev below that of the fresh surface. Such a change might be ascribed to the action of small amounts of H_2 which appear to be generated in the oil diffusion pumps as a result of a gradual breakdown of the hydrocarbons. Helium and argon glow discharges caused only small changes in ϕ in all cases, with helium producing more consistent results, usually yielding an average decrease of about 0.2 ev. Argon treatment gave somewhat more scattered values which, averaged over a large number of runs, gave zero change.

Of the electropositive gases, H_2 gave much more consistent data than N_2 with an average decrease of about 0.3 ev. With N_2 bombardment there appeared to be a tendency to lower ϕ by 0.1 to 0.2 ev. In a few cases larger negative changes were noted.

The most consistent results were obtained for the glow discharge treatments in the electronegative gases,

[8] H. B. Michaelson, J. Appl. Phys. 21, 536 (1950).

O_2 and freon-12[9]. After bombardment the work functions increased in all cases by amounts between 1.0 and 1.4 ev. It was found impossible to restore surfaces to their initial condition either by electron bombardment or by glow discharges in helium or argon. In contrast to the gases previously discussed (He, A, H_2, N_2) which may form surface films by physical adsorption, it appears that oxygen and the discharge decomposition products of freon-12 may be absorbed chemically, possibly producing thin films of Ag_2O, AgF, and AgCl which would permanently contaminate the surface even though no visible surface changes could be detected. The magnetron filament did not seem to alter its surface in these processes since the emission remained constant after flashing.

Water vapor at 20-mm Hg pressure was admitted to the system and caused a decrease of the work function of clean silver by only 0.2 ev in one case and by a negligible amount in the other. This is in agreement with Oatley's findings[6] for alternate layers of hydrogen and oxygen.

All of the results discussed above and presented in Figs. 2 and 3 are in general accord with those of most previous observers.[5] The oxygen data show good quantitative agreement.

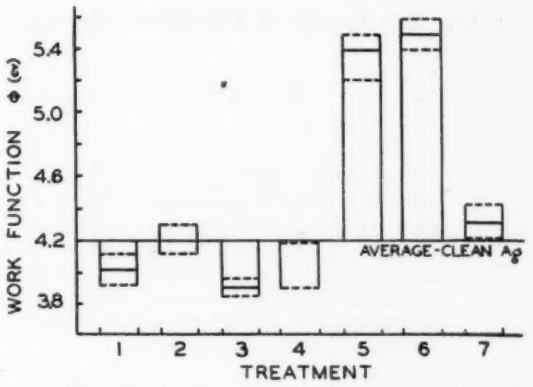


FIG. 3. Work functions of five silver surfaces. The glow discharge treatments in the respective gases are indicated by numbers along the abscissa: 1 in He; 2 in A; 3 in H_2; 4 in N_2; 5 in O_2; 6 in freon; 7 in H_2O. In the case of H_2O the surface was only exposed to the vapor, not its glow discharge. The dashed lines indicate the spread of the experimentally observed values.

The authors wish to acknowledge their indebtedness to Professor L. B. Loeb who initially suggested the problem and gave encouragement throughout its progress. Thanks are also due to Mr. S. W. Bale and his staff for the fine mechanical work, and to Mr. Robert Greiner for his excellent glass blowing. The help of Mr. R. W. Kotter during the early data taking is also appreciated.

[9] Kinetic Chemicals, Inc., 10th and Market Streets, Wilmington, Delaware, Technical Paper No. 6 on freon-12 (Underwriter's Report), and Technical Paper No. 8. Freon-12 has the formula CCl_2F_2, and it is most probable that its decomposition products consisting of fluorine and chlorine compounds were responsible for the observed effects.

JOURNAL OF APPLIED PHYSICS VOLUME 24, NUMBER 4 APRIL, 1953

On the Zeros of Polynomials and the Degree of Stability of Linear Systems*

J. FRANK KOENIG
Office of Basic Instrumentation, National Bureau of Standards, Washington, D. C.
(Received January 12, 1953)

A method is presented for finding the zeros of any nth degree real or complex polynomial as its coefficients (or their parameters) are varied. Several new theorems are presented. The method is based on new theorems, some concerning determinants, and on known theorems. It is applied to characteristic equations of linear physical systems with or without feedback. Root trajectories in the complex plane can be visualized without plotting them from "root trajectory diagrams" which show curve families for real or complex roots of a polynomial equation in a plane having two coefficients of the polynomial equation as coordinates. The diagrams can be used in the converse manner to determine the coefficient values corresponding to a desired dominant complex pair or real root.

1. INTRODUCTION

IN the design of physical systems used for measurement and control and in other mechanical, electrical, hydraulic, or pneumatic devices, it is important to know whether the effect of transient disturbances will die out. If the proposed system can be defined by a linear differential equation with constant coefficients, one method of meeting the stability requirements is to form the corresponding characteristic equation and adjust its coefficients so that all its roots lie in the left half of the complex plane. Furthermore, the greater the distance between the imaginary axis and the complex conjugate pair nearest it, the greater the degree of stability, provided all other roots lie to the left of and well beyond this dominant pair. Its location can be specified by two coordinates, say an angle and a length, or two lengths. The necessary adjustment of the coefficients can be made by utilizing the relationships among them, or among the physical parameters of which they consist, and the real and imaginary parts of the roots.

These relationships are stated here as theorems, all but one using various determinants which can be formed from the coefficients of the real polynomial constituting the left side of the characteristic equation. These theorems have also a more general usefulness in locating all the zeros of any nth degree polynomial with either real or complex coefficients. This paper presents new theorems and applies these in combination with known theorems to the problem of locating all the roots of a characteristic equation when coefficients or parameters are varied. The method presented can be termed the "root trajectory" method, since the diagrams shown will provide data for plotting the path (trajectory) which a root takes in the complex plane as the result of such variation. The designer can visualize these trajectories, and they are not shown herein. For simplicity it is desirable to consider the simultaneous variation of only two or three parameters. The diagrams can be used in the converse manner to determine the parameter values corresponding to a desired dominant complex pair or real root.

A graphical method for obtaining all roots of a characteristic equation, with the coefficients assigned numerical values, was developed by Leonhard.[1,2] Evans[3] developed a graphical method called the root locus method, using a mechanical device called a Spirule to obtain loci of all roots of the characteristic equation.

2. KNOWN DETERMINANT CRITERIA

In this section two known theorems will be stated which are expressible in terms of determinants. The theorems will be applied later as part of the root trajectory method.

If a polynomial has real coefficients, then by application of the criterion of Routh[4,5] and Hurwitz[6] it can be determined whether or not all its zeros have negative real parts. The Hurwitz theorem is as follows:

Theorem 1. The roots of the nth degree polynomial equation with real coefficients,

$$p^n+a_1p^{n-1}+a_2p^{n-2}+\cdots+a_n=0, \qquad (1)$$

have negative real parts if and only if all determinants

$$H_k=\begin{vmatrix} a_1 & a_3 & a_5 & \cdots & a_{2k-1} \\ 1 & a_2 & a_4 & \cdots & a_{2k-2} \\ 0 & a_1 & a_3 & \cdots & a_{2k-3} \\ 0 & 1 & a_2 & \cdots & a_{2k-4} \\ 0 & 0 & a_1 & \cdots & a_{2k-5} \\ \cdot & \cdot & \cdot & \cdots & \cdot \\ 0 & 0 & 0 & \cdots & a_k \end{vmatrix}$$

are positive for $k=1, 2, 3, \cdots, n$. The elements a_i are zero if the index i is larger than n or negative.

If a polynomial has complex coefficients, then by

* Part of the work reported here was completed in the Department of Electrical Engineering, University of Illinois. The part completed in the Office of Basic Instrumentation, National Bureau of Standards, was conducted under a cooperative program of Basic Instrumentation Research and Development sponsored by the U. S. Office of Naval Research, U. S. Air Force, and the U. S. Atomic Energy Commission.

[1] A. Leonhard, Arch Elektrotech. 39, 100 (1948).
[2] A. Tustin, Editor, *Automatic and Manual Control* (Academic Press, Inc., New York, 1951), A. Leonhard, pp. 25–35.
[3] W. R. Evans, Trans. Am. Inst. Elec. Engrs. 69, 66 (1950).
[4] E. J. Routh, *Stability of a Given State of Motion* (MacMillan and Company, Ltd., London, 1877).
[5] E. J. Routh, *Dynamics of a System of Rigid Bodies* (MacMillan and Company, Ltd., London, 1905), Part 2, Chap. 6.
[6] A. Hurwitz, Math. Ann. 46, 273 (1895).

application of the criterion of Bilharz[7] and Frank[8] it can be determined whether or not all its zeros have negative real parts. The Bilharz-Frank theorem states:

Theorem 2. All roots of the complex polynomial equation

$$p^n+(A_1+jB_1)p^{n-1}+(A_2+jB_2)p^{n-2}+\cdots+(A_n+jB_n)=0 \quad (1a)$$

have negative real parts if and only if all determinants

$$\Delta_1=A_1$$

and

$$\Delta_k=\begin{vmatrix} A_1 & A_3 & A_5 & \cdots & A_{2k-1} & -B_2 & -B_4 & \cdots & -B_{2k-2} \\ 1 & A_2 & A_4 & \cdots & A_{2k-2} & -B_1 & -B_3 & \cdots & -B_{2k-3} \\ \cdot & \cdot & \cdot & \cdots & \cdot & \cdot & \cdot & \cdots & \cdot \\ \cdot & \cdot & \cdot & \cdots & \cdot & \cdot & \cdot & \cdots & \cdot \\ 0 & 0 & 0 & \cdots & A_k & 0 & 0 & \cdots & -B_{k-1} \\ 0 & B_2 & B_4 & \cdots & B_{2k-2} & A_1 & A_3 & \cdots & A_{2k-3} \\ 0 & B_1 & B_3 & \cdots & B_{2k-3} & 1 & A_2 & \cdots & A_{2k-4} \\ \cdot & \cdot & \cdot & \cdots & \cdot & \cdot & \cdot & \cdots & \cdot \\ \cdot & \cdot & \cdot & \cdots & \cdot & \cdot & \cdot & \cdots & \cdot \\ 0 & 0 & 0 & \cdots & B_k & 0 & 0 & \cdots & A_{k-1} \end{vmatrix}$$

are positive; $k=2, 3, \cdots, n$, with $A_i=B_i=0$ for $i>n$. The Bilharz-Frank criterion is given by Marden.[9]

3. ROOT LOCI IN THE COMPLEX PLANE

3.1 Sectors

One may suppose Eq. (1) to be the characteristic equation of some linear physical system. Assume its roots lie to the left of the imaginary axis of p. To find whether they also lie within the left-hand sector $\pi-2\beta$ in Fig. 1, where $0\leq\beta<\pi/2$, substitute the transformation $p=ze^{-i\beta}$ into Eq. (1), obtaining an equation in the complex variable z. This transformation relates the z plane to the p plane. If the p axes and z axes are considered to be coincident, then this substitution rotates the z roots counterclockwise by an angle β with respect to the p roots. Suppose the z roots are then made coincident with the p roots by rotating the z plane clockwise through β with respect to the p axes. Subsequent discussion relates to this latter configuration, shown in Fig. 1.

The z equation is

$$z^n+(A_1+jB_1)z^{n-1}+\cdots+(A_n+jB_n)=0, \quad (2)$$

in which the A_i and B_i are real. Suppose the Bilharz-Frank criterion is applied successfully (all determinants positive) to Eq. (2); then all roots of Eq. (2) have negative real parts, and lie above the tilted $Im(z)$ axis. Suppose further that the Hurwitz criterion has been applied successfully (all determinants positive) to Eq. (1); then all z roots are located in the left half of the p plane. If all z roots are located in the left half of the p plane and also above the $Im(z)$ axis, then all p roots are located within the left-hand sector $\pi-2\beta$ since p roots and z roots now coincide, and complex p roots always occur in conjugate pairs. This proves:

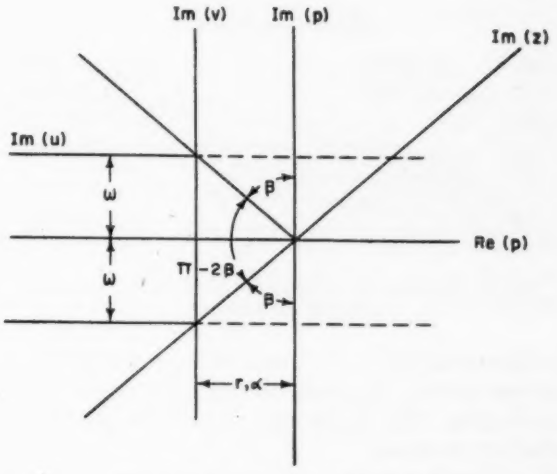


FIG. 1. Superposed complex planes.

Theorem 3. The roots of the polynomial equation (1) are located within the sector $\pi-2\beta$ (Fig. 1), if the Routh-Hurwitz criterion is applied successfully (all Hurwitz determinants positive) to Eq. (1), and the Bilharz-Frank criterion is applied successfully (all determinants positive) to the complex polynomial equation (2) obtained by substituting $p=ze^{-i\beta}$ into Eq. (1). β satisfies $0\leq\beta<\pi/2$.

The information provided by Theorem 3 has been obtained by graphical methods by Leonhard[10] and Vazsonyi.[11]

Suppose that all p roots are within the sector $\pi-2\beta$. Then suppose the coefficient values of Eq. (1) continuously varied so that one complex pair of roots moves onto the *boundary* of the sector. Under this condition Theorem 3 no longer applies. However, the following theorem does apply:

Theorem 4. If the complex polynomial equation (2) has at least one pure imaginary root, then the last Bilharz-Frank determinant Δ_n is zero.

The proof is given in Sec. 6. We assume for the present that the converse of Theorem 4 is true. This matter will be discussed in Sec. 6.

Now continue the variation of the coefficients of Eq. (1) until some roots lie outside the sector $\pi-2\beta$, at least one pair is *on* the boundary, and some roots may perhaps remain within the sector. Under these conditions Theorem 4 still applies, and the last Bilharz-Frank determinant Δ_n is zero. Since Δ_n relates β to all the p coefficients, the determinant equation $\Delta_n=0$ can be plotted in Cartesian coordinates, by varying two of the coefficients of Eq. (1) and holding all other coefficients constant. For any fixed value of β each set of values of

[7] H. Bilharz, Z. Angew. Math. Mech. **24**, 77 (1944).
[8] E. Frank, Bull. Am. Math. Soc. **54**, 144, 890 (1946).
[9] M. Marden, *The Geometry of the Zeros* (American Mathematical Society, New York, 1949), p. 140.
[10] A. Leonhard, reference 2.
[11] A. Vazsonyi, J. Appl. Phys. **20**, 863 (1949).

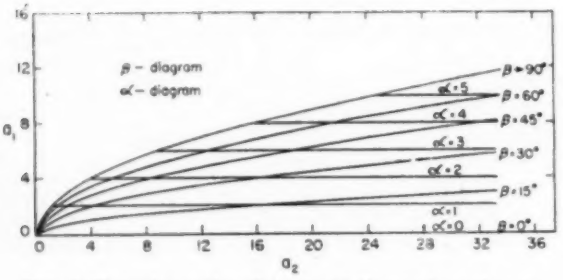


FIG. 2. Damping-ratios diagram (β-diagram); real-parts diagram (α-diagram).

the two variables a_i, for which at least one pure imaginary root of Eq. (2) exists, corresponds to a point on the $\Delta_n=0$ curve. This Cartesian plot will be called a root trajectory diagram.

Consider the characteristic equation of a second-order system

$$p^2+a_1p+a_2=0, \tag{3}$$

where a_1 and a_2 are real and adjustable. It is often desirable to relate these coefficients in such a way as to provide a specified damping ratio $|\alpha/\omega|$, where $-\alpha\pm j\omega$ is the complex pair of roots. In Fig. 1, $\beta=\tan^{-1}|\alpha/\omega|$. The trajectory diagram for such complex roots may be found as follows. Substitute $p=ze^{-j\beta}$ into Eq. (3), which gives

$$z^2+a_1e^{j\beta}+a_2e^{j2\beta}=0. \tag{4}$$

Equation (4) can be written

$$z^2+a_1(\cos\beta+j\sin\beta)z+a_2(\cos2\beta+j\sin2\beta)=0, \tag{5}$$

which is of the form of Eq. (2). The last Bilharz-Frank determinant is

$$\Delta_2=\begin{vmatrix} a_1\cos\beta & 0 & -a_2\sin2\beta \\ 1 & a_2\cos2\beta & -a_1\sin\beta \\ 0 & a_2\sin2\beta & a_1\cos\beta \end{vmatrix}.$$

Set $\Delta_2=0$. The plot appears in Fig. 2 as the β-diagram. All points on a given β-parabola correspond to a complex conjugate pair of roots of Eq. (3) lying on the sides of the corresponding sector in Fig. 1.

3.2 Translated Half-Planes

To see whether the roots of Eq. (1) have real parts more negative than $-\alpha$ with $\alpha\geq0$, substitute $p=v-\alpha$ into Eq. (1). Then suppose the $Im(v)$ axis translated to the left, Fig. 1, so that v roots become coincident with p roots. The Cypkin[12]-Bromberg[12]-Grossman[13]-Bothwell[14,15] theorem can be stated:

Theorem 5. The real parts of all roots of a real polynomial equation (1) in p will be more negative than $-\alpha$, α real, if and only if upon substitution of $p=v-\alpha$ all Hurwitz v determinants are positive.

This theorem is an extension of the Routh-Hurwitz criterion. A theorem by Orlando,[16] Duncan,[17] Frazer,[17] and Sponder[18] is as follows:

Theorem 6. If $\alpha=0$ in Theorem 5, and the v equation has a pure imaginary pair of roots, then the next to last Hurwitz v determinant H_{n-1} is zero.

The converse of this theorem is true except for two special cases, discussed by Orlando. Sponder described one of the two special cases. The following theorem is an extension of Theorem 6.

Theorem 7. The v equation obtained in Theorem 5 has a pure imaginary pair of roots if and only if the next to the last Hurwitz v determinant H_{n-1} is zero, except for the two special cases discussed by Orlando.

Theorem 7 is essentially a theorem developed by Cypkin, Bromberg, and Bothwell. However, these authors apparently were not aware of the two special cases. It will be shown later that the two special cases are of little practical significance.

Suppose all roots of Eq. (1) do have real parts more negative than $-\alpha$, $\alpha\geq0$. Then allow the coefficient values of Eq. (1) to vary continuously so that one complex pair of roots moves onto the half-plane boundary (the imaginary v axis, Fig. 1). Under this condition Theorem 7 applies. Cypkin, Bromberg, and Bothwell used Theorem 7 to locate the real parts of the one complex pair having the least negative real part. Their theorem will be generalized in order to obtain the real parts of *all* complex roots of Eq. (1). The generalization is as follows:

Theorem 8. The real parts of *all* complex roots of any nth degree real polynomial equation (1) can be obtained by substituting $p=v-\alpha$, α real, into Eq. (1) and equating the next to last Hurwitz v determinant H_{n-1} to zero, except in the two special cases mentioned above.

To prove Theorem 8 it is only necessary to show that Theorem 7 can be thus generalized. This follows from the fact that in the equation $H_{n-1}=0$, with v coefficients assigned numerical values, α is the only quantity left with an unassigned value. Hence, if there are k complex p pairs, one may assign k values of α to satisfy Theorem 7.

Justification will now be given for stating that the two special cases are of little practical significance. For any nth degree real polynomial equation (1), $H_{n-1}=0$ is satisfied if and only if either (a) at least one pair of roots is pure imaginary (b) at least one real root is the negative of another real root, or (c) at least one complex root is the negative of another complex root. The conditions (b) and (c) are the two special cases for which the converse of Theorem 7 is not true and are the exceptions in Theorem 8.

The equation $H_{n-1}=0$ is a function of all coefficients

[12] Ya. Z. Cypkin and P. V. Bromberg, Izvestia Akad. Nauk SSSR, Otdelenie Tekhnicheskikh Nauk, 1163 (1945). (In Russian.)
[13] K. H. Grossman, Schweiz. Arch. angew. Wiss. Tech. **14**, 242 (1948).
[14] F. E. Bothwell, Proc. Inst. Radio Engrs. 1345 (1950).
[15] F. E. Bothwell, Trans. Am. Inst. Elec. Engrs. **69**, 1430 (1950).
[16] L. Orlando, Math. Ann. **71**, 233 (1912). (In Italian.)
[17] W. J. Duncan and R. A. Frazer, Proc. Roy. Soc. (London) **A124**, 624 (1929).
[18] E. Sponder, Schweiz. Arch. angew. Wiss. Tech. **16**, 93 (1950); Translation in NACA TM 1348 (August, 1952).

(or parameters) of the real polynomial equation (1). Suppose these coefficients are fixed in such a way that one root pair is purely imaginary. Now vary two of the coefficients or parameters continuously. Experience shows that this can generally be done so as to keep this root pair purely imaginary, and satisfy the equation $H_{n-1}=0$. (In general, it is not possible to maintain one pair purely imaginary with the continuous variation of only *one* coefficient or parameter.) With a purely imaginary pair, the two varying coefficients need control only one root since its conjugate follows automatically. The two real roots of case (b), however, have no such automatic connection. They cannot, in general, be controlled in such a way that one is the negative of the other as the two coefficients are varied to satisfy $H_{n-1}=0$. The same type of argument holds for case (c). Thus the plot of $H_{n-1}=0$ in terms of the continuous variation of any two coefficients a_x and a_y is a continuous curve for case (a), but is a set of isolated points for cases (b) and (c).

Applying Theorem 8 to the quadratic example, substitute $p=v-\alpha$ into Eq. (3). The transformed equation is

$$v^2+(-2\alpha+a_1)v+(\alpha^2-a_1\alpha+a_2)=0. \tag{6}$$

Equation (6) is a real polynomial of the form

$$v^2+c_1v+c_2=0. \tag{7}$$

For Eq. (7) the next to last Hurwitz determinant H_{n-1} is $H_1=c_1$. Set $H_1=c_1=0$. This equation is plotted in Figs. 2 and 3 as the family of straight lines parallel to the a_2 axis. For all points on the $\alpha=2$ horizontal, Eq. (3) has two complex roots with real part equal to -2. This family of straight lines is called an α-diagram.

The half-plane translation was also used by Cypkin, Bromberg, and Bothwell to locate also the one *real* root with the least negative real part. Their theorem is:

Theorem 9. If $p=v-r$ ($r>0$ and $-r$ is the least negative real root) is substituted into a real polynomial equation (1) in p, the constant term of the resulting v equation is zero.

This can be generalized as follows:

Theorem 10. *All* the real roots of a real or complex polynomial equation of the nth degree, Eqs. (1) or (1a), can be obtained by substituting $p=v-r$ and equating to zero the constant term of the v equation.

The proof of this theorem follows the reasoning used to prove Theorem 8.

Applying Theorem 10 to the quadratic example, set c_2 of Eq. (7) equal to zero, after replacing α in Eq. (6) by r, with $r>0$. This corresponds to zero-value roots of Eq. (7) and to real p roots equal to $-r$. The equation $c_2=0$ is plotted in Fig. 3 as a family of tangent lines, an r diagram. All points on the $r=2$ tangent correspond to a real root of Eq. (3) equal to -2. The point at which the $r=1$ tangent intersects the $r=5$ tangent corresponds to real roots of Eq. (3) equal to -1 and -5.

3.3 Infinite Strips

To see whether the complex roots of Eq. (1) have imaginary parts whose absolute values are less than ω with $\omega>0$, make the transformation $p=-j(u+\omega)$. Then, to make u roots coincident with p roots, suppose the $Im(u)$ axis rotated 90° counterclockwise and translated upward by ω to the position shown in Fig. 1. If all Bilharz-Frank u determinants are positive, then all u roots will lie below the shifted $Im(u)$ axis. Since any complex p roots will occur in conjugate pairs, they will lie no farther than 2ω below it.

Suppose that all p roots of Eq. (1) do have imaginary parts whose absolute values are less than ω with $\omega>0$. All p roots are located within the infinite strip, Fig. 1. Vary the coefficients of Eq. (1) simultaneously so that at least one complex pair of p roots moves onto the boundary of the infinite strip. Under this condition the last Bilharz-Frank u determinant Δ_n is zero, by Theorem 4. Continuing the quadratic example, substitute $p=-j(u+\omega)$ into Eq. (3). The transformed equation is

$$u^2+(A_1+jB_1)u+(A_2+jB_2)=0, \tag{8}$$

or

$$u^2+(2\omega+ja_1)u+(\omega^2-a_2+j\omega a_1)=0. \tag{9}$$

The last Bilharz-Frank determinant is

$$\Delta_2=\begin{vmatrix} 2\omega & 0 & -a_1 \\ 1 & (\omega^2-a_2) & -a_1 \\ 0 & a_1 & 2\omega \end{vmatrix}. \tag{10}$$

Suppose that for some value of ω, Eq. (8) has a pure imaginary root; then $\Delta_2=0$ and Eq. (3) must have a pair of complex conjugate roots with imaginary parts $\pm\omega$. In Figs. 3 and 4 the equation $\Delta_2=0$ is plotted as a family of parabolas. All points on the $\omega=3$ parabola correspond to a complex conjugate pair of roots of Eq. (3) with imaginary parts ± 3. The diagram in Fig. 4 can be called an ω-diagram. We have assumed here that the converse of Theorem 4 is true. This matter will be discussed in Sec. 6.

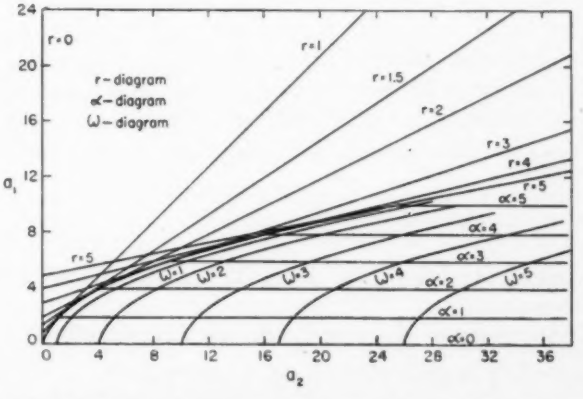


FIG. 3. Real-roots diagram (r-diagram); real-parts diagram (α-diagram); imaginary-parts diagram (ω-diagram).

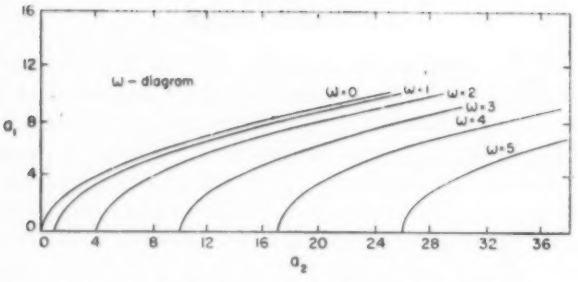


FIG. 4. Imaginary-parts diagram (ω-diagram).

4. SUPERPOSITION OF DIAGRAMS

Consider Fig. 2 which shows the superposition of the β-diagram and the α-diagram. A complex pair of roots of Eq. (3) is obtained from the intersection of one α-curve and one β-curve. Alternatively, complex conjugate pairs of roots of Eq. (3) will occur where the α-horizontals intersect the ω-parabolas, e.g., the point at which the $\alpha=2$ line intersects the $\omega=3$ parabola corresponds to the pair $-2\pm j3$.

In Fig. 3 all straight lines of the r diagram are tangent to the $\omega=0$ curve. The r diagram and α-diagram intersect on the $\omega=0$ parabola, all points of which correspond to real, equal, and negative roots. For example, the $r=1$ tangent intersects the $\alpha=1$ horizontal on the $\omega=0$ parabola; the roots are $-1, -1$. All points on and above the $\omega=0$ parabola correspond to real roots. All points on the a_2 axis correspond to purely imaginary roots. The $\beta\rightarrow 90°$ and $\omega=0$ parabolas are identical.

5. APPLICATION TO PHYSICAL SYSTEMS

In another paper[19] the root trajectory method has been used to determine the degree of stability of a third-order servomechanism. The diagrams are there plotted in terms of the values of the physical parameters. The method can be applied to any nth order system; its coefficients may be functions of several parameters, say m of them. Two plots in m-dimensional space, one for real and one for complex roots, would show the effect of simultaneous variation of m parameters. For ready visualization of the root trajectories it is preferable to display only two or three parameters at a time, in Cartesian coordinates. The labor involved in obtaining root trajectory diagrams naturally increases with the degree of the characteristic equation.

To provide satisfactory transient behavior of a linear physical system, the designer might specify that its characteristic equation have a dominant pair of complex conjugate roots. "Dominant" here means having the least negative real part and the smallest damping ratio, $\tan\beta=|\alpha/\omega|$; it is assumed that there are no other roots nearby and that no roots have positive real parts. A *real* root is dominant under these same two assumptions, when it is less negative than the real parts of all other roots. In the method shown, one diagram (i.e., superposition of any two of the α-, ω-, and β-diagrams) gives the parameter values corresponding to the desired pair of dominant complex roots; the other diagram (r diagram) gives the parameter values corresponding to a dominant real root. The designer concerned with degree of stability is interested only in a limited range of dominant roots, and it is then sufficient to plot only those areas of the root trajectory diagrams for roots lying in the region between the imaginary axis and the vicinity of the desired dominant roots. With these diagrams the designer can determine whether a point exists in the diagrams corresponding to a sufficiently dominant root or pair. The root trajectory method was applied to the quadratic system in another paper.[20]

For any nth degree polynomial equation, root trajectory diagrams can be plotted if some of the coefficients are negative. However, for a stable physical system it is necessary, but not sufficient, that all coefficients of Eq. (1) be of the same sign.

6. PROOF OF THEOREM 4

Theorem 4 was stated in Sec. 3: If the complex polynomial equation (2) has at least one purely imaginary root, then the last Bilharz-Frank determinant Δ_n is zero. A form of the last Bilharz-Frank determinant Δ_n given by Marden[21] is

$$\Delta_n=\begin{vmatrix} A_1 & -B_2 & -A_3 & B_4 & A_5, & \cdots, & (-1)^{n+1} & A_{2n-1} \\ 1 & -B_1 & -A_2 & B_3 & A_4, & \cdots, & (-1)^{n+1} & A_{2n-2} \\ 0 & A_1 & -B_2 & -A_3 & B_4, & \cdots, & (-1)^{n+1} & B_{2n-2} \\ 0 & 1 & -B_1 & -A_2 & B_3, & \cdots, & (-1)^{n+1} & B_{2n-3} \\ 0 & 0 & A_1 & -B_2 & -A_3, & \cdots, & (-1)^{n} & A_{2n-3} \\ 0 & 0 & 1 & -B_1 & -A_2, & \cdots, & (-1)^{n} & A_{2n-4} \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdots & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdots & \cdot & \cdot \end{vmatrix}, \tag{12}$$

in which $A_j=B_j=0$ for $j>n$. Δ_n has $2n-1$ rows and columns.

Substitute $p=j\omega$, ω real, into Eq. (2). We obtain

$$(j\omega)^n+(A_1+jB_1)(j\omega)^{n-1}+(A_2+jB_2)(j\omega)^{n-2}+\cdots+(A_n+jB_n)=0. \tag{13}$$

Equation (13) is composed of a real part and an imaginary part. Assume that Eq. (13) is satisfied for some

[19] J. F. Koenig, Proc. Natl. Electronics Conf. 7, 24 (1951).

[20] J. F. Koenig, "Additions to the Stability Theory and Design of Servomechanisms," American Institute of Electrical Engineers Graduate Student Paper, Great Lakes District Meeting, May 17–19, 1951.

[21] M. Marden, reference 7, p. 139.

value of ω. Therefore the real and imaginary parts of Eq. (13) are individually equal to zero. The two equations are of the following form:

$$-\omega^n - B_1\omega^{n-1} + A_2\omega^{n-2} + B_3\omega^{n-3} - A_4\omega^{n-4} - B_5\omega^{n-5} + \cdots = 0, \quad (14)$$

$$-A_1\omega^{n-1} - B_2\omega^{n-2} + A_3\omega^{n-3} + B_4\omega^{n-4} - A_5\omega^{n-5} - B_6\omega^{n-6} + \cdots = 0. \quad (15)$$

In Eq. (14) the last term is $-A_n$ if n is even, and B_n if n is odd. In Eq. (15) the last term is B_n if n is even, and $-A_n$ if n is odd. We now make use of Sylvester's[22] determinant:

$$D(f,g) = \begin{vmatrix} a_0 & a_1 & a_2 & \cdots & a_n & 0 & \cdots & \\ 0 & a_0 & a_1 & \cdots & \cdot & a_n & 0 & \cdot \\ \cdot & \cdot & \cdot & \cdots & \cdot & \cdot & \cdot & \cdot \\ 0 & \cdots & a_0 & a_1 & \cdots & \cdot & a_n & \\ b_0 & b_1 & \cdots & b_m & 0 & \cdots & \cdot & \\ 0 & b_0 & b_1 & \cdot & \cdot & b_m & 0 & \cdots \\ \cdot & \cdot & \cdot & \cdots & \cdot & \cdot & \cdots & \cdot \\ 0 & 0 & 0 & \cdots & b_0 & b_1 & \cdot & b_m \end{vmatrix} \begin{matrix} \\ \\ \\ \\ \\ \\ \\ \end{mat

cients of Eq. (2) are fixed in such a way that one root of Eq. (2) is pure imaginary. Now vary two of the coefficients or parameters of Eq. (2) continuously. Experience shows that this can generally be done so as to keep this root purely imaginary, and satisfy the equation $\Delta_n=0$. (In general, it is not possible to maintain one root pure imaginary with the continuous variation of only *one* coefficient or parameter.) Thus the plot of $\Delta_n=0$ in terms of two coefficients or parameters a_x and a_y of Eq. (2) is a continuous curve for the case corresponding to a purely imaginary root of Eq. (2), and the special cases correspond to a set of isolated points, by the argument previously given for H_{n-1}, and are of little practical significance.

ACKNOWLEDGMENTS

The author thanks Dr. G. H. Fett, Professor of Electrical Engineering, University of Illinois, for helpful criticisms during the course of the work. Thanks are expressed to Dr. H. L. Mason and Dr. N. A. Wiegmann, National Bureau of Standards, for helpful discussions regarding preparation of the manuscript.

JOURNAL OF APPLIED PHYSICS VOLUME 24, NUMBER 4 APRIL, 1953

Approximate Elastic Spectrum of Acoustic Waves in AgCl from X-Ray Scattering*

H. Cole
Massachusetts Institute of Technology, Cambridge, Massachusetts
(Received December 31, 1952)

The temperature diffuse scattering of x-rays from a large single crystal of AgCl has been measured at room temperature using a Geiger counter spectrometer and crystal monochromated CuK-α radiation. Measurements were made throughout the *hk*0 plane in reciprocal space, and in certain other regions. The data are analyzed on the basis of the theory developed by Laval and James, in which temperature vibrations are expressed in terms of traveling elastic waves. An assumed set of vibrational frequencies is used for the optical branch. Dispersion curves for waves in the acoustical branch traveling in the [100], [110], and [111] directions are obtained. The velocities at infinite wavelength are shown to agree fairly well with those given by the static elastic constants. An approximate frequency spectrum is then calculated, and suggestions are made as to how the effects of the optical and acoustical branches may be separated.

I. INTRODUCTION

Olmer,[1] Cole and Warren,[2] and Curien[3] have obtained the frequencies of a large number of elastic waves of different wavelengths, propagating through large single crystals having cubic lattices and containing essentially one type of atom, by measuring the diffuse scattering of x-rays from aluminum, beta-brass, and iron. An ordered cubic crystal containing two types of atoms, such as AgCl, differs from the cases already studied only in the fact that there are six frequencies associated with each elastic wave vector rather than three; a "longitudinal" and two "transverse" in the optical branch, and a "longitudinal" and two "transverse" in the acoustical branch. The same experimental techniques may be used in the case of AgCl to determine many of these frequencies as a function of wave vector and type of oscillation, provided some additional method is employed to separate the contributions to the measured intensities of modes in the two frequency branches.

From Laval,[4] we may write for the intensity per unit cell of the first order temperature diffuse scattering from AgCl at any position P in reciprocal space, defined by $\mathbf{hb}_1+\mathbf{kb}_2+\mathbf{lb}_3+\mathbf{g}$:

$$I_{eu}=4\left(\frac{\sin\theta}{\lambda}\right)^2 e^{-2M}\left[\sum_{i=1}^{3}\frac{(f_1\pm\gamma_i f_2)^2}{m_1+m_2\gamma_i^2}\frac{\bar{E}_i\cos^2(\mathbf{S},\mathbf{e}_i)}{\nu_i^2}+\sum_{i'=1}^{3}\frac{(\Gamma_{i'}f_1\pm f_2)^2}{m_1\Gamma_{i'}^2+m_2}\frac{\bar{E}_{i'}\cos^2(\mathbf{S},\mathbf{e}_{i'})}{\nu_{i'}^2}\right]_s \quad (1)$$

where *hkl* are the Miller indices for the nearest fcc Bragg reflection, $\mathbf{b}_1\mathbf{b}_2\mathbf{b}_3$ are the reciprocal lattice vectors, and the scattered intensity is due only to the traveling elastic wave with wave vector $\pm\mathbf{g}$ where $|\mathbf{g}|=1/\Lambda$ and Λ is the wavelength of the elastic wave. 2θ is the scattering angle, λ is the wavelength of the monochromatic x-ray beam, and e^{-2M} is the Debye temperature factor. Actually there is a Debye factor for each atom type, however, for AgCl the two factors are so nearly identical at room temperature that a single common factor may be used. $\sum_{i=1}^{3}$ and $\sum_{i'=1}^{3}$ are the sums over the three acoustical and the three optical modes associated with each elastic wave. f_1 and f_2 are the atomic scattering factors for Ag^+ and Cl^- with a Hönl dispersion correction. m_1 and m_2 are, respectively the masses of Ag^+ and Cl^-. γ_i is the ratio of the mean square ampli-

* Research sponsored by the U. S. Office of Naval Research under Contract N5ori-07832.

[1] P. Olmer, Acta Cryst. **1**, 57, 1948; Bull. Soc. Franc. Minéral **71**, 144 (1948).
[2] H. Cole and B. E. Warren, J. Appl. Phys. **23**, 335 (1952).
[3] H. Curien, Acta Cryst. **5**, 393 (1952).
[4] J. Laval, Bull. Soc. Franc. Minéral **64**, 1 (1941).

tude of oscillation of the Cl^- ion to that of the Ag^+ ion, for the acoustical modes associated with this wave vector. $\Gamma_{i'}$ is the ratio of the mean square amplitude of Ag^+ to that of Cl^- for the optical modes. $\bar{E}_i$ and $\bar{E}_{i'}$ are the average energies of the acoustical and the optical modes. These may be written as the average energy of a linear oscillator of frequency ν_i or $\nu_{i'}$, in thermal equilibrium at the temperature of the crystal, for the acoustical modes $\bar{E}_i \rightarrow kT$. $\mathbf{S}/\Lambda$ is a vector from the origin of the reciprocal lattice to the point of measurement, so that $\mathbf{S}=\lambda(\mathbf{hb}_1+\mathbf{kb}_2+\mathbf{lb}_3+\mathbf{g})$. e_i and $e_{i'}$ are unit vectors pointing in the direction of oscillation of the mode, and ν_i and $\nu_{i'}$ are the frequencies. $(\mathbf{S}, \mathbf{e}_i)$ is the angle between the vectors $\mathbf{S}$ and $\mathbf{e}_i$. The plus signs are used when P is in the first Brillouin zone centered on an even (hkl even) reflection; the negative signs are used when in an odd (hkl odd) zone.

By choosing appropriate positions in reciprocal space, the cosine squared terms may be made to go to zero for all but one mode. For example, along a line from the origin of reciprocal space in the [100] direction, all transverse oscillations are at right angles to $\mathbf{S}$, and the longitudinal oscillations are parallel to $\mathbf{S}$. Equation (1) then reduces to

$$I_{eu}[100]=4\left(\frac{\sin\theta}{\lambda}\right)^2 e^{-2M}\left[\frac{kT(f_1\pm\gamma_{l,\,ac}f_2)^2}{m_1+m_2\gamma^2_{l,\,ac}}\frac{1}{\nu^2_{l,\,ac}}+\frac{\bar{E}_{l,\,op}(\Gamma_{l,\,op}f_1\pm f_2)^2}{m_1\Gamma^2_{l,\,op}+m_2}\frac{1}{\nu^2_{l,\,op}}\right], \quad (2)$$

where l, ac and l, op stand for longitudinal acoustic and longitudinal optical.

For a crystal containing only one atom type, the second term in the bracket vanishes, as do the gammas, and so an intensity measurement at any point along such a line gives directly the frequency of the longitudinal mode traveling in the [100] direction, with a wavelength equal to the reciprocal of the distance from the point of measurement to the nearest Bragg reflection. Using such geometrical considerations, Olmer, Cole and Warren, and Curien obtained dispersion curves for each mode of the elastic waves traveling in simple directions and from these dispersion curves synthesized approximate frequency distribution curves. However, to obtain such dispersion curves for elastic waves in AgCl, the effects of the optical and acoustical frequencies for each mode type have to be separated, and some means found for evaluating the gamma-factors, or atomic amplitude ratios.

Wooster's[5] group at the Cavendish laboratories have used diffuse intensity measurements to evaluate the elastic constants of crystals of the AgCl type. The measurements were taken close to a Bragg reflection, since only the long waves or short $\mathbf{g}$ vectors were of interest for this work. For short-wave vectors the acoustical frequencies approach zero, while the optical frequencies remain high. For these long waves each atom type oscillates with the same amplitude in the acoustical vibrations, and so in Eq. (1) the second term in the bracket may be neglected with respect to the first term and γ set equal to unity in the remaining term. Under these conditions, Eq. (1) reduces to that used in the previous investigations.

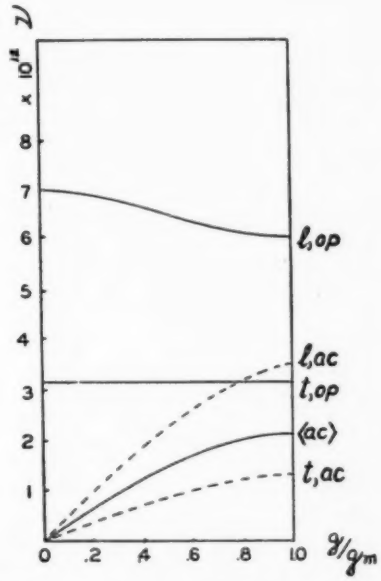


FIG. 1. Trial frequency spectrum for AgCl, used to correct for the contribution of the optical branch.

The gap between the acoustical and optical branches is largely a matter of the difference in the masses of the two atom types. With a large mass difference, the contribution of the optical branch remains small and often negligible for all wave vectors. The contribution was assumed to be small in AgCl, where Ag has approximately three times the mass of Cl. Actually a trial frequency spectrum was constructed, which is shown by Fig. 1, using the known elastic constants, reststrahlen frequency, Kellermann's[6] theoretical treatment of NaCl, and the theoretical results for a one-dimensional diatomic chain. These assumed optical frequencies were used in the evaluation of the acoustical frequencies from the experimental data, and if they are reasonably correct, their effect is quite small.

The gamma-factors are functions of the mode, and are largely a function of the mode frequency as is the mode energy. For a one-dimensional diatomic chain with only nearest neighbor interaction, the amplitude ratios can be calculated explicitly,[7] and Fig. 2 is a plot of these ratios for a one-dimensional diatomic chain with masses equal to those of Ag^+ and Cl^-. It was assumed that to a first approximation these values could be used for waves propagating in the AgCl lattice in the direc-

[5] G. N. Ramachandran and W. A. Wooster, Acta Cryst. 4, 335 (1951); 4, 431 (1951).

[6] E. W. Kellermann, Phil. Trans. A238, 513 (1940).

[7] L. Brillouin, *Wave Propagation in Periodic Structures* (McGraw-Hill Book Company Inc., New York, 1946), p. 151.

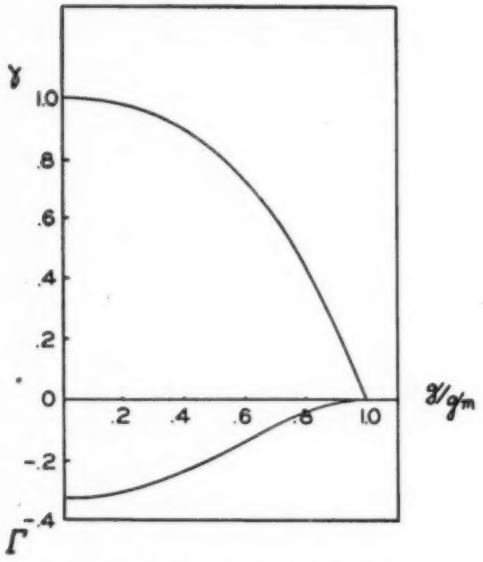


FIG. 2. Atomic amplitude ratios used to interpret the scattering data for AgCl.

tions for which dispersion curves were obtained. For zero wave vector $\gamma=1.0$, and it falls off to zero at g(max), indicating that the Cl^- ions come to rest in the acoustical branch at g(max). In the optical branch the Ag^+ ions are at rest at g(max), but increase their amplitude of oscillation as $\mathbf{g}$ decreases, until at $\mathbf{g}=0$ their amplitude is about $\frac{1}{3}$ the amplitude of the Cl^- ions. This corresponds to the condition existing at the time of excitation of the reststrahlen frequency. Γ is negative, since in the optical modes the atoms oscillate "out of phase." Fairly complete theoretical discussions of lattice vibrations and their effect in scattering x-rays, are given by Laval,[4] Brillouin,[7] Born,[8] and James.[9]

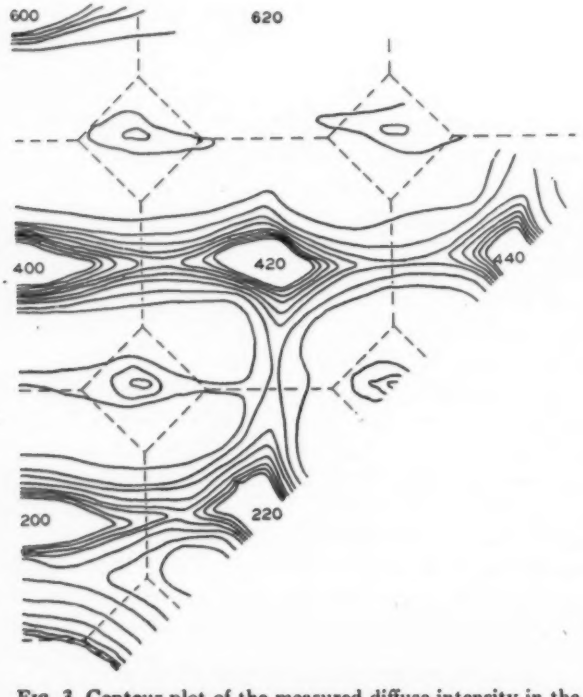


FIG. 3. Contour plot of the measured diffuse intensity in the $hk0$ plane of the reciprocal lattice of AgCl, $T=20°C$.

II. EXPERIMENTAL PROCEDURE

The experimental set-up was the same as that used by Cole and Warren.[2] X-rays from a full-wave rectified copper target tube, run at 35 kv and 16 ma, were monochromated by a quartz crystal $(10\bar{1}1)$ bent to a radius of 45 cm and set to reflect CuK-α. The beam was focused on the face of the crystal. This was mounted in a goniometer so that any desired position in reciprocal space could be reached, while maintaining equal angles of incidence and diffraction, to keep the absorption correction angular independent. A nickel-aluminum balanced filter was used to eliminate the half wavelength reflected by the quartz crystal. The scattered radiation was measured by a Geiger counter using 4-minute

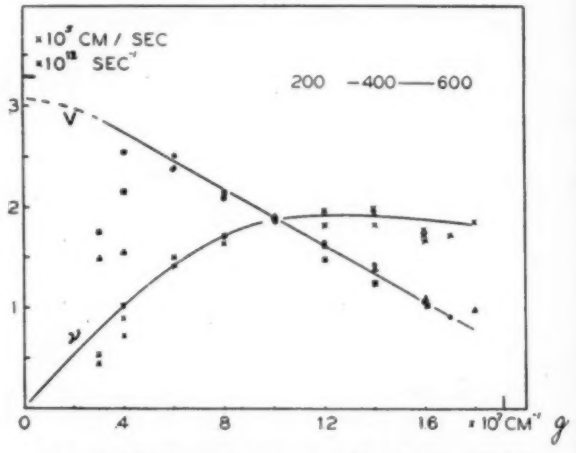


FIG. 4. Velocity and frequency of longitudinal [100] acoustical waves in AgCl as a function of $g=1/\Lambda$.

counting intervals. The volume element averaged over in reciprocal space at each reading was that defined by a 2.5° divergence in the primary beam and counter slits 2.0 mm square, 10 cm from the crystal. The count with the aluminum foil in the beam was about twice the general background in the regions of lowest intensity. All data were taken at room temperature.

The crystal was a large (2.5 cm×2.5 cm×1.5 cm) single crystal of AgCl cut from an ingot grown from the melt by Harshaw Chemical Company. It was loaned to us by Dr. C. D. West of the Polaroid Corporation. All data were taken using a face cut parallel to the (110) crystalline planes. This face was scanned by back reflection Laue photographs and the spots were quite sharp over the whole face. A thin layer of free crystalline silver some few cells deep was detected on the surface by its diffraction pattern. The layer was oriented with

[8] M. Born, Rept. Prog. Phys. (Physical Society, London) 9, 294 (1942–1943).

[9] R. W. James, *The Optical Principles of the Diffraction of X-Rays* (G. Bell and Sons, London, 1948), Chap. 5.

respect to the base crystal, but it was considered to be too thin to affect the intensity measurements. All counts were converted to absolute units by comparison with the high angle scattering from a thick block of paraffin.

III. RESULTS

A contour plot of the recorded difference count in the *hk*0 plane of the reciprocal lattice of AgCl is shown by Fig. 3. Since the Compton modified scattering varies slowly, the general shape of the contours is due to the temperature diffuse scattering. The traces of the Brillouin zones in this plane are indicated by the dotted lines. The "peaks" at the forbidden reflection positions are actually cross sections of temperature diffuse arms running through these positions at right angles to the *hk*0 plane, connecting the crystalline reflections above and below the plane of the paper. Along the [100] axis the reflections are disk shaped, with the plane of the disk perpendicular to [100]. The reflections become more nearly spherical as one moves away from the [100] axis.

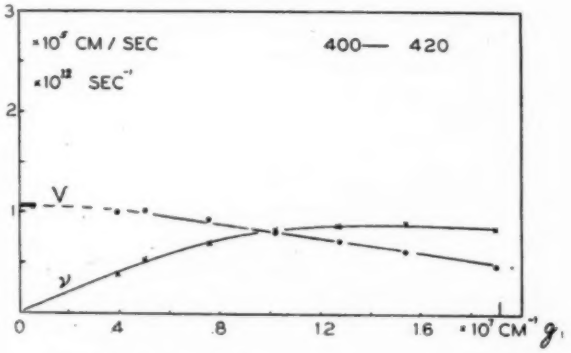


FIG. 5. Velocity and frequency of transverse [100] acoustical waves in AgCl as a function of $g=1/\Lambda$.

In the central regions of Fig. 3, the difference count was about 4×64 counts in 4 minutes. The first contour represents a count of 6×64, and each succeeding contour represents an increase of 2×64 counts. The highest count represented on the plot is 20×64. The contours agree very closely with those predicted by Jahn's[10] method, using the known elastic constants of AgCl.

Data were taken in several other regions outside of the *hk*0 plane to obtain the desired dispersion curves. The measured intensities have to be corrected for air scattering, Compton modified scattering, and higher order temperature diffuse scattering. In the case of AgCl, the second-order temperature scattering was not negligible and actually amounted to as much as 50 percent of the first-order scattering at large angle. The correction for second-order scattering from a cubic lattice containing two atom types is no more difficult to calculate than that for a lattice containing only one atom type, since the acoustical and optical contributions

[10] H. Jahn, Proc. Roy. Soc. (London) **A179**, 320 (1942).

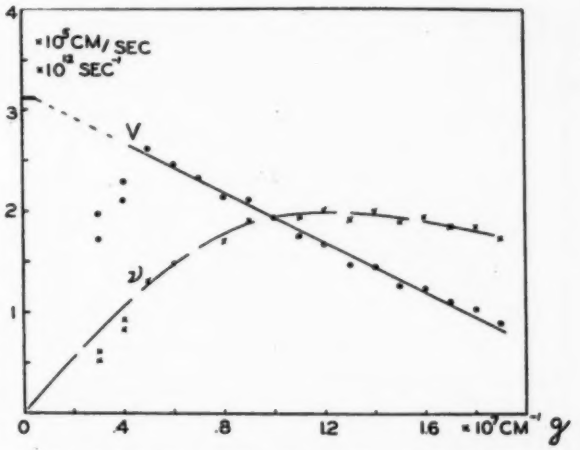


FIG. 6. Velocity and frequency of longitudinal [110] acoustical waves in AgCl as a function of $g=1/\Lambda$.

can be calculated separately. The second-order scattering correction was computed using the trial frequency spectrum and gamma-factors shown in Figs. 1 and 2.

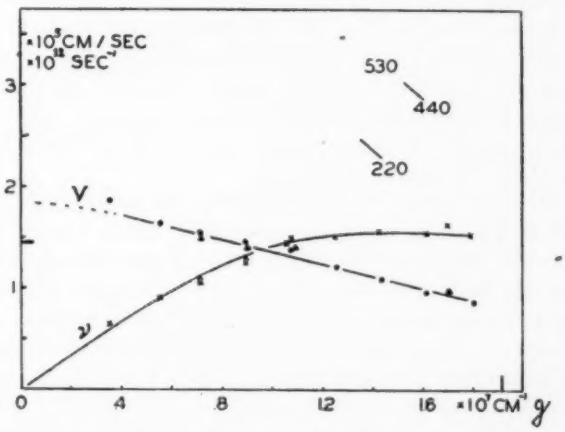


FIG. 7. Velocity and frequency of the t_1 transverse [110] acoustical waves in AgCl as a function of $g=1/\Lambda$.

Figures 4 through 10 are plots of the phase velocity and frequency *versus* $1/\Lambda$ for each of the three acoustical modes of waves traveling in the [100], [110], and [111]

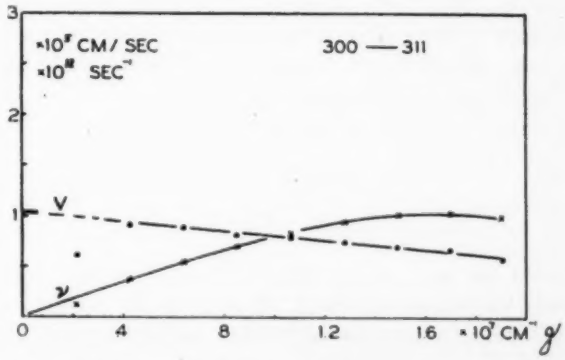


FIG. 8. Velocity and frequency of the t_2 transverse [110] acoustical waves in AgCl as a function of $g=1/\Lambda$.

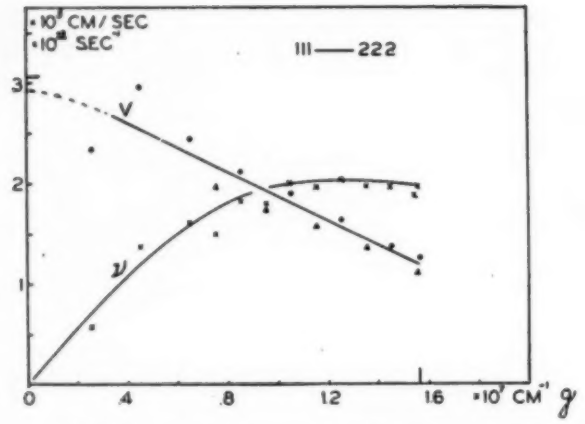


FIG. 9. Velocity and frequency of longitudinal [111] acoustical waves in AgCl as a function of $g=1/\Lambda$.

directions. These curves were obtained by applying Eq. (1) to the experimentally measured intensities along the paths indicated in the plots, after correcting for Compton modified and second-order temperature scattering, and using the assumed atomic amplitude ratios and optical frequencies shown in Figs. 1 and 2.

The limiting velocities as **g** goes to zero should agree with those obtained from elastic constant measurements. A comparison of the limiting velocities given by the intercepts of these curves and those obtained by Arenberg[11] is shown in Table I.

The agreement is not as good as might be desired. The longitudinal modes were expected to be low, as in Cole and Warren's[2] measurements on beta-brass. It is believed that this effect is due to the primary beam's being insufficiently monochromatic, which tends to spread out the crystalline reflections along radial directions.† This would give too high intensity readings at small **g** for longitudinal modes and so would lower the velocity curve. The sudden drop of the experimental points at small **g**, shown in some of the curves, is an indication that the counter had begun to pick up some of the Bragg reflection. There seems to be no obvious reason why the $t_1(110)$ and $t(111)$ modes are high. If the x-ray data were correct, c_{12} would have to be lowered from about 0.36×10^{12} to about 0.26×10^{12} dynes/cm². However, the Bulk modulus $(c_{11}+2c_{12})/3$ has been measured by a number of investigators, and the results are in good agreement. It must be concluded that there are small errors in the x-ray data.

Figure 11 is a plot of the quantity

$$\frac{1}{\rho}\frac{dn}{d\nu}=\frac{\nu^2}{V^2U}$$

versus ν for each of the mode types whose dispersion curves have been obtained. Here dn is the number of waves with frequencies between ν and $\nu+d\nu$, ρ is the volume of the crystal in cm³, V is the phase velocity, and U is the group velocity given by the slope of the dispersion curve. These curves have been roughed in, and for calculations reference should be made to the actual dispersion curves. The curves of Fig. 11 may be put together by Houston's[12] method or any other means of interpolation to obtain an over-all frequency distribution. The small sketch represents an approximate result from these data with the assumed optical contributions also shown.

IV. DISCUSSION

As a result of the rather unexpected behavior of the terms $(f_1\pm\gamma f_2)^2$ and $(\Gamma f_1\pm f_2)^2$ in Eqs. (1) and (2), the dispersion curves of Figs. 4–10 are probably better than would be expected on the basis of all the approximations that have been necessary. On the basis of the assumed forms for γ and Γ, $(\Gamma f_1\pm f_2)^2$ approaches zero, since Γ is negative, as one approaches a crystalline reflection in an even zone, and so the optical contribution vanishes, no matter what the actual frequencies are. Since all but one of the dispersion curves were obtained from measurements in an even zone, the dispersion curves of the acoustical modes are probably fairly good.

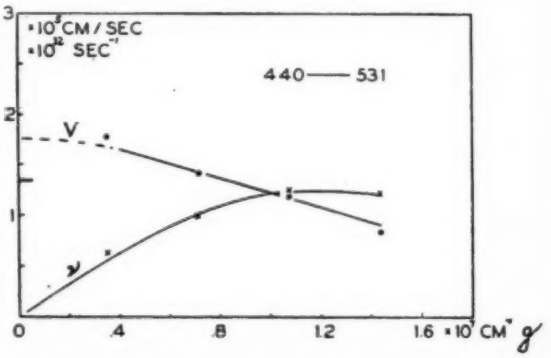


FIG. 10. Velocity and frequency of transverse [111] acoustical waves in AgCl as a function of $g=1/\Lambda$.

This behavior of the "structure factors," $(f_1\pm\gamma f_2)^2$ and $(\Gamma f_1\pm f_2)^2$ may afford a method for separating the effects of the optical and acoustical modes. In an even (*hkl* even) zone, the contribution of the optical modes is a maximum out at the edge of the zone, and goes to zero as one approaches the Bragg reflection, while the acoustical modes, since γ is positive, peak in the center and die off toward the edge. On the other hand, in an odd zone (*hkl* odd), both contributions would tend to peak at the center. Therefore, by taking intensity measurements at the same **g** position in an odd and even zone, it may be possible by solving pairs of simultaneous equations, to obtain both the acoustical and optical frequencies experimentally on the basis of the assumed atomic amplitude ratios.

[11] D. L. Arenberg, J. Appl. Phys. **21**, 941 (1950).

† This could also result from the size and shape of the volume element in reciprocal space.

[12] W. V. Houston, Revs. Modern Phys. **20**, 161 (1948).

Theoretically, it would also be possible to separate the effects of the optical and acoustical branches by taking measurements at various temperatures. The only factor which changes greatly with T is the average energy of the mode, and under certain conditions, this has a quite different temperature dependence for the mode types in the two frequency branches. For example, if ν acoustical is much smaller than ν optical, it might be possible to find a temperature range over which the energy of the acoustical mode still varies as kT, while the optical mode has already reached its zero-point energy. By taking measurements at several temperatures in this range, and extrapolating to $T=0$, the contribution of the two modes may be separated. However, for the actual frequencies involved in AgCl, measurements would have to be made at liquid air temperature or lower, and although the total scattering remains fairly intense even at these low temperatures, it is doubtful if the small shift of one component with respect to the other could be detected with present day experimental techniques.

TABLE I. Comparison of limiting velocities for AgCl obtained by ultrasonic and x-ray methods.

Mode	Elastic theory	Ultrasonic	X-ray scattering
$l(100)$	$\rho v^2=c_{11}$	3.29×10^5 cm/sec	3.0×10^5 cm/sec
$l(110)$	$2\rho v^2=c_{11}+c_{12}+2c_{44}$	3.13×10^5 cm/sec	2.8×10^5 cm/sec
$l(111)$	$3\rho v^2=c_{11}+2c_{12}+4c_{44}$	3.07×10^5 cm/sec	2.8×10^5 cm/sec
$t(100)$	$\rho v^2=c_{44}$	1.06×10^5 cm/sec	1.06×10^5 cm/sec
$t_2(110)$	$\rho v^2=c_{44}$	1.06×10^5 cm/sec	1.06×10^5 cm/sec
$t_1(110)$	$2\rho v^2=c_{11}-c_{12}$	1.466×10^5 cm/sec	1.75×10^5 cm/sec
$t(111)$	$3\rho v^2=c_{11}-c_{12}+c_{44}$	1.344×10^5 cm/sec	1.75×10^5 cm/sec

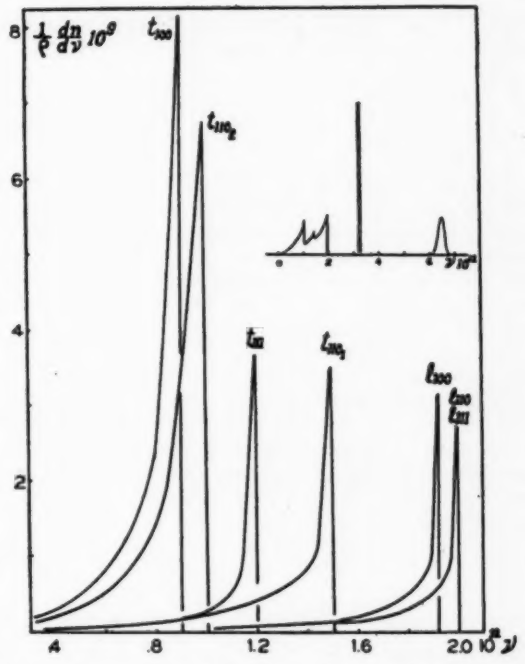


FIG. 11. Frequency distribution curves for longitudinal and transverse acoustical waves in AgCl for the [100], [110], and [111] directions. Small insert: approximate frequency distribution for AgCl averaged over all directions, including the assumed optical frequencies.

The droop in some of the frequency curves of Figs. 4–10 at large $\mathbf{g}$ may be due to a real second nearest neighbor interaction, which would give such a droop, or it may be an indication that not enough optical contribution was subtracted out, since the optical contribution rises with increasing $\mathbf{g}$. If this is the case, the assumed optical frequencies, especially for the longitudinal optical (l, op) branch, may be a little high or do not drop off enough for high $\mathbf{g}$ values.

Constructing an over-all frequency distribution from the few dispersion curves given here has the disadvantage of accentuating the peculiarities of the distributions in these few directions. However most of the peculiarities seem to be in the transverse modes, as the three longitudinal modes seem to make more or less identical contributions to the total distribution. Actually it is not necessary to go through the dispersion curves to get the frequency distribution. Since most intensity measurements can to a good approximation be made to represent a single mode frequency, intensity measurements may be converted directly to frequencies, the frequencies tabulated and then grouped into a distribution. The distribution calculated on this basis has not as yet been evaluated because of the tedious second-order temperature scattering corrections for points off of simple lattice directions.

The author is indebted to Professor B. E. Warren for his continuous assistance, and to Professors J. C. Slater, N. H. Frank, and B. Averbach for helpful discussions and suggestions. The author also wishes to thank Dr. C. D. West of the Polaroid Corporation for the loan of the AgCl crystal.

The Limiting Negative Pressure of Mercury in Pyrex Glass

LYMAN J. BRIGGS
National Bureau of Standards, Washington, D. C.
(Received December 31, 1952)

It is generally considered that mercury does not "wet" glass, as evidenced by the convex surface of the mercury meniscus in a manometer. If this were true under all conditions, the development of a negative pressure in a mercury column would be impossible. But *pure* mercury in a clean evacuated *degassed* tube shows a *flat* meniscus. In other words, when the tube is degassed, mercury does adhere to the glass wall.

A U-tube manometer of 5-mm bore with one leg sealed off was evacuated and degassed at 500°C with a mercury-vapor pump, and then filled with pure mercury by distillation. The vertical height of the mercury column from the meniscus in the open tube to the top of the closed leg was 52 cm. When the manometer was again evacuated, the mercury column remained suspended from the top of the tube, withstanding a negative pressure of two-thirds of an atmosphere.

In extending the measurements, short mercurial thermometers were first used. Each thermometer was cemented to the spinner of a variable-speed motor, with the spin axis bisecting the distance from the end of the bulb to the meniscus. The maximum centrifugal stress on the free capillary column was consequently at the spin axis. Measurement of nine thermometers gave negative pressures ranging from 2 to 17 bars.

Finally, freshly drawn capillary tubes with fine bores were evacuated, degassed by heating, sealed off and filled by breaking one end under pure mercury. The negative pressure which these mercury columns could withstand under centrifugal stress increased rapidly as the degassing improved. The highest negative pressure observed for mercury was 425 bars at 28°C. In this case, the furnace temperature during degassing was raised to the point where the capillary tube started to deform.

INTRODUCTION

THE familiar convex meniscus of mercury in a clean U-tube manometer is generally interpreted to mean that mercury does not "wet" the glass wall. In fact, "standard" tables are now used in precise barometry to correct the observed height of the column for the curvature of the meniscus, calculated on the assumption that the adhesion between mercury and glass is zero. If this were literally true under all conditions, then it would, of course, be impossible to develop a negative pressure in a mercury-glass system.

But physicists working with diffusion pumps often obtain what is known as a "sticking" vacuum in a McLeod gage, i.e., the mercury in the measuring capillary on release does not immediately follow the balancing column down, but hangs from the top of the capillary and develops a *negative* pressure of 1 or 2 cm of mercury before letting go. This "sticking" vacuum is more easily realized if the McLeod gage is carefully torched during the evacuation.

Wichers[1] and his colleagues at the National Bureau of Standards have found that base metal impurities in mercury can be reduced to 10^{-7} by the simple expedient of bubbling air through the mercury and then drawing off the purified metal from below to separate it from the oxidized metals floating on the surface. The criterion of this remarkably high purity is simply a clean mirror-like surface. The addition of a base metal, such as zinc, copper, lead, or tin, to this pure mercury, in a concentration of only 1 part in 10 000 000 again fouls the surface.

These investigators also found that when a clean evacuated ampule was partially filled with pure dry mercury, the resulting meniscus was *flat*. This clearly indicates adhesion of the mercury to the glass wall. Even when *dry* air was admitted, the meniscus remained flat. Sealed tubes of this kind, prepared 10 years ago, still show this characteristic. I have repeated this experiment with larger ampules of Pyrex glass, 1 cm internal diameter. The tubes were evacuated with a diffusion pump, degassed at about 500°C, partially filled with mercury by distillation and sealed off. One tube was then opened to the air of the laboratory, when the meniscus assumed the characteristic convex form. The meniscus in the other evacuated tube was *flat*, out to the very wall of the tube. In fact, an examination of the edge of the meniscus under a magnifying glass gives the impression that the mercury is trying to climb the wall.

[1] Wichers, Chem. Eng. News 20, September, 1942.

MEASUREMENTS WITH MERCURY MANOMETERS

The above observations led me to think that under suitable conditions mercury may withstand a considerable negative pressure. The first experiments were made with manometers.

A U-tube manometer of Pyrex glass with one leg sealed off was evacuated and degassed at about 500°C with a mercury-vapor pump. With the pump still operating it was then filled with mercury by distillation. The length of the mercury column from the sealed end to the meniscus in the open leg was 52 cm. The bore of the tube was 5 mm.

After filling, this manometer was mounted vertically and the open leg was evacuated. But nothing happened. The mercury column remained hanging from the top of the closed leg. This was so startling that I mistrusted the pump until it had been tested against another manometer.

This experiment demonstrated conclusively that the negative pressure which mercury in glass was capable of sustaining exceeded two-thirds of an atmosphere. Only by vertical jarring was the column finally released, when the mercury assumed the same level in both legs

of the tube. But once released, I could never get this column to hang up again.

This experiment was repeated with 3 additional manometers, similarly filled,* but with a column length of 26 cm. All of them withstood a negative pressure of one-third of an atmosphere. One of them "healed" itself after being jarred down and continued to withstand the original negative pressure.[2]

The difficulties involved in baking out and handling long glass tubes filled with mercury made it impractical to try to extend these direct hydrostatic measurements. Accordingly, I turned to a modification of the centrifugal method used in investigating the limiting negative pressure of water[3] and other liquids.[4]

THE CENTRIFUGAL METHOD

In the centrifugal method, the liquid being studied is held in a capillary tube mounted on a horizontal spinner attached to the vertical axis of a variable-speed motor. The capillary tube is so mounted that the midpoint of the liquid column is intersected by the spin axis.[4] One-half of the liquid column thus pulls against the other half. The maximum stress at the mid-point is

$$\tfrac{1}{2}\rho r^2\omega^2 \text{ dynes/cm}^2 = \text{bars}/10^6,$$

where ρ = density of liquid, r = length of one-half column in cm, and ω = angular speed in radians per second at the instant the column breaks.

EXPERIMENTS USING MERCURIAL THERMOMETERS

Centrifugal measurements were first made on short mercurial thermometers. The thermometer under test was cemented to the spinner in such a position that the distance from the spin axis to the end of the bulb was slightly greater than the distance to the meniscus. The spinner was then carefully balanced. When the column broke under test, it was thrown outward into the expansion bulb at the end of the capillary tube.

The measurements were not very satisfying. First, the gas pressure in the thermometer was unknown in most instances. It was assumed to be one atmosphere. This enters as a correction, since the gas pressure must first be overcome by the centrifugal force before any *negative* pressure is developed. This uncertainty was eliminated in tests on 3 evacuated thermometers.† Second, the column sometimes failed by a segment flying off the outer end, instead of rupturing at the midpoint. Consequently, observations at the midpoint used with other liquids[3,4] could not be relied upon. It was necessary to increase the speed in increments, stopping the motor after each increase in speed to determine whether the column was still intact. Observations were sometimes continued on a shortened column after remeasuring the length.

The negative pressures developed in 8 mercurial thermometers, in the order observed, are as follows: 14, 11, 4, 17, 7, 4.4, 2.2, 2 bars (1 bar = 0.986 atmos). The wide variation is probably attributable to differences in cleanliness and dryness of the thermometer capillaries. The bore was generally elliptical in cross section, with a major diameter of about 0.2 mm and a minor diameter of about 0.11 mm. The measurements were made at room temperature.

TABLE I. Limiting negative pressure of mercury in Pyrex glass.

Tube no.	P_n (bars)	Treatment
9	46	Torched
10	10	Not torched
11	7	Not torched
12	47	Furnace, 200°C
	57	Furnace, 200°C
13	277	Furnace, 450°–500°C
	323	Furnace, 450°–500°C
14	338	Furnace, 450°–500°C
15	425	Furnace, yield point
16	383	Furnace, yield point

MEASUREMENTS WITH EVACUATED CAPILLARY TUBES

The measurements given above suggested that much higher negative pressures might be obtained if special attention were given to evacuating and drying the capillary tubes. The following technique was employed. A fine capillary about 20 cm long was drawn on the end of a Pyrex tube and sealed off at one end, the other end being left attached to the larger tube for evacuation. About 2 cm from its free end, the capillary was bent sharply back upon itself to form a hairpin loop. The capillary was then evacuated and sealed off about 12 cm above the loop. The loop was next broken off under pure dry mercury, completely filling the tube; and finally, a small amount of mercury was driven off and the end sealed off, leaving a free space at the end of the capillary into which the column could move when it ruptured under centrifugal stress.

The results are given in Table I. Comparison of tube 9 with 10 and 11 shows the importance of heating the tube during evacuation, even if the capillaries are freshly drawn. The remaining tubes were accordingly heated in a tubular electric furnace during evacuation.

It will be noted that two values are given for tubes 12 and 13. In each case, as the speed of the motor was increased, a short segment of mercury first tore off from the outer end of the column and moved to the sealed-off end of the capillary. This gave the first value. The shortened column was then remeasured and the speed increased in successive steps until the mercury column broke at its intersection with the spin axis, which led to the second higher value.

* I am indebted to Leonardo Testa, National Bureau of Standards, for constructing and filling all of these manometers.

[2] This experiment was demonstrated before the National Academy of Sciences in Washington, D. C. on April 30, 1952.

[3] L. J. Briggs, J. Appl. Phys. **7**, 721 (1950).

[4] L. J. Briggs, J. Chem. Phys. **19**, 970 (1951).

† These thermometers were kindly supplied by Dr. R. D. Thompson, Taylor Instrument Company.

The breaking off of the free end of the column led me to think that even the momentary exposure of the tube to the atmosphere during the final sealing-off process might modify the measurements. The technique was accordingly changed to avoid this exposure. The evacuated tube was clamped to the side of a tripod, with the hairpin bend partly submerged in the mercury-filled porcelain crucible under the tripod. Breaking the hairpin filled the tube, leaving the lower end still immersed. The tube was then sealed off *above* the tripod with the aid of a weak spiral spring which pulled the capillary away when the fusion was complete. This left a free evacuated space at the end of the mercury column.

Tubes 15 and 16 were prepared in this way, and withstood higher negative pressures than any of the earlier tubes. The temperature of the furnace was raised to the point where the vertical capillary distorted slightly under the asymmetric load of the hairpin bend. One end of tube 16 broke off under the high centrifugal stress, ending the measurement.

The highest value obtained for the negative pressure of mercury exceeded 425 bars at 27°C (tube No. 15). The stress was increased in 16 steps, starting with a negative pressure of 120 bars. After each increment of speed the motor was stopped for inspection of the column. The length of the free leg from the spin axis to the meniscus was 57 mm. The break occurred within 3 mm of the spin axis. The bore of the capillary was about 0.14 mm.

This value for mercury, 425 bars, is believed to be the highest direct measurement of negative pressure so far reported for any liquid. The next highest, 318 bars, is for chloroform.[4]

JOURNAL OF APPLIED PHYSICS VOLUME 24, NUMBER 4 APRIL, 1953

On Unsteady Flow of Compressible Viscous Fluids*

G. LUDFORD, *University of Maryland, College Park, Maryland*
H. POLACHEK,† *Naval Ordnance Laboratory, White Oak, Silver Spring, Maryland*
AND
R. J. SEEGER, *National Science Foundation, Washington, D. C.*
(Received January 5, 1953)

In order to solve the nonlinear equations of fluid dynamics, classical hydrodynamics introduced the concept of perfect incompressible fluids, *viz.*, the perfect inviscid fluid and the perfect viscous fluid. The practical importance of considering compressibility nowadays has resulted in two new approaches from two distinct points of view. In the one case, compressibility is included in the discussion of a perfect inviscid fluid. As is well known, however, compressibile effects include shock-wave phenomena, which are regarded in this instance as mathematical discontinuities. The goal then is to study the basic equations, allowing discontinuous solutions. The other approach retains the viscous terms together with the compressible ones. It is based upon the possibility of small coefficients producing large changes in the behavior of the solutions. The present paper is written from this point of view. The consequent mathematical difficulties of solving the equations are avoided by the use of a particle model of the continuum; the authors are indebted to Professor R. von Mises of Harvard University for this idea. The resulting equations are put into finite difference form for solution on a digital calculator.

THE PARTICLE MODEL

THE equations of unsteady motion in one dimension of a perfect viscous compressible fluid may be written

$$\rho\frac{Du}{Dt}=-\frac{\partial p}{\partial x}+\frac{\partial \sigma}{\partial x}, \tag{1}$$

$$\frac{D\rho}{Dt}=-\rho\frac{\partial u}{\partial x}, \tag{2}$$

$$\sigma\frac{\partial u}{\partial x}=\rho T\frac{DS}{Dt}, \tag{3}$$

$$p=\rho\bar{R}T;\quad \bar{R}=C_v(\kappa-1), \tag{4}$$

where

$$\sigma=-\frac{4}{3}\mu\frac{\partial u}{\partial x}, \tag{5}$$

$$S=C_v\log\frac{p}{\rho^{\kappa}} \tag{6}$$

are the viscous stress and entropy per unit mass, respectively. Here C_v is the specific heat at constant volume, μ is the coefficient of viscosity, and κ the gas exponent; these will all be assumed constant. Heat conduction can also be included in the model, but is omitted here since it was not considered in the computations given later.

The energy Eq. (3) may be transformed by use of Eqs. (4) and (6) to

$$\sigma\frac{\partial u}{\partial x}=\frac{p}{\kappa-1}\left[\frac{1}{p}\frac{Dp}{Dt}-\frac{\kappa}{\rho}\frac{D\rho}{Dt}\right];$$

* This work was carried out at the Naval Ordnance Laboratory under the sponsorship of the Mathematics Branch of the U. S. Office of Naval Research.
† Now at David Taylor Model Basin, Corderock, Maryland.

thus, by use of equation of continuity (2) it becomes

$$\frac{Dp}{Dt}=\frac{\partial u}{\partial x}[(\kappa-1)\sigma-\kappa p], \tag{7}$$

which does not contain ρ.

The motion of the gas, which may be assumed to take place within a tube of unit cross section, will be approximated by the motion of a system of particles, each of mass m, which move on a fixed straight line. Thus, as each particle moves, it is supposed to represent that mass m of the gas which initially had the particle at its center. Neighboring particles are subject to interactions which approximate the viscous and pressure forces in the original gas. Thus if x_n is the x coordinate of the nth particle, and a dot represents differentiation with respect to time t, the viscous interaction between the (nth) and the ($n+1$)th particles will be taken as

$$\sigma_{n+\frac{1}{2}}=\frac{4}{3}\mu\frac{\dot{x}_{n+1}-\dot{x}_n}{x_{n+1}-x_n} \tag{8}$$

[see Eq. (5)]. Equation (1) will then be replaced by

$$m\ddot{x}_n=-(p_{n+\frac{1}{2}}-p_{n-\frac{1}{2}})+(\sigma_{n+\frac{1}{2}}-\sigma_{n-\frac{1}{2}}), \tag{9}$$

where the pressure interaction $p_{n+\frac{1}{2}}$ between the (nth) and ($n+1$)th particles will satisfy an equation

$$\dot{p}_{n+\frac{1}{2}}=\frac{\dot{x}_{n+1}-\dot{x}_n}{x_{n+1}-x_n}[(\kappa-1)\sigma_{n+\frac{1}{2}}-\kappa p_{n+\frac{1}{2}}], \tag{10}$$

which is the equivalent of Eq. (7).

Equations (8)–(10) are sufficient to characterize the particle model, if it is assumed that the constants μ, κ are those of the original gas. Thus the equation of continuity (2) is automatically satisfied by the nature of the model.

Consider now the special case in which the original gas is contained in a stationary closed tube of length l and has initially a uniform velocity V, a pressure p_0, and a density ρ_0. This may be realized in practice by the sudden arrest of the tube when both it and its contents are moving with a uniform velocity V. Then in the corresponding particle model a number $N+2$ particles, each of mass $m=\rho l/(N+1)$, are used, of which the end ones $n=0$ and $n=N+1$ are kept fixed for all time t, and the others $n=1, \cdots, N$ have initially a velocity V.

The equations of the particle motion may be made nondimensional by the transformations

$$\begin{cases} p=p_0P=\dfrac{1}{\kappa}\rho_0a_0^2P; & \sigma=\dfrac{\mu}{l}V\Sigma^* \\ x=lX; & t=\dfrac{lT^*}{V}; \quad a_0^2=\dfrac{\kappa p_0}{\rho_0}. \end{cases} \tag{11}$$

The resulting equations will contain as parameters the Mach number $M=V/a_0$, and the Reynolds number $R=Vl\rho_0/\mu$; the boundary and initial conditions will be absolute. In order to make the equations depend on only one parameter the transformation

$$\begin{cases} T^*=M\left(\dfrac{\kappa}{N+1}\right)^{\frac{1}{2}}T \\ \Sigma^*=\dfrac{4}{3M}\left(\dfrac{N+1}{\kappa}\right)^{\frac{1}{2}}\Sigma \end{cases} \tag{12}$$

is made. This gives in place of Eqs. (8)–(10),

$$\begin{cases} \Sigma_{n+\frac{1}{2}}=\dfrac{X'_{n+1}-X'_n}{X_{n+1}-X_n} \\ X_n''=-(P_{n+\frac{1}{2}}-P_{n-\frac{1}{2}})+A(\Sigma_{n+\frac{1}{2}}-\Sigma_{n-\frac{1}{2}}) \\ P'_{n+\frac{1}{2}}=\dfrac{X'_{n+1}-X'_n}{X_{n+1}-X_n}[(\kappa-1)A\Sigma_{n+\frac{1}{2}}-\kappa P_{n+\frac{1}{2}}], \end{cases} \tag{13}$$

where dashes denote differentiation with respect to T, and

$$A=\frac{4M}{3R}[\kappa(N+1)]^{\frac{1}{2}}. \tag{14}$$

The boundary and initial conditions are not absolute now; indeed,

$$\begin{aligned} &T=0: \quad X_n=\frac{n}{N+1}, \quad X_n'=\left(\frac{\kappa}{N+1}\right)^{\frac{1}{2}}M=B; \\ &\qquad\qquad\qquad\qquad\qquad\qquad n=1, \cdots, N \\ &\qquad P_{n+\frac{1}{2}}=1 \qquad\qquad\qquad\qquad n=0, \cdots, N \\ &T\geq 0: \quad X_0=0, \qquad X_{N+1}=1. \end{aligned} \tag{15}$$

To recapitulate: in order to use a model of $N+2$ particles to approximate the motion under discussion, the parameters A and B are first computed from the known initial state and constants of the gas by means of Eqs. (14) and (15). Then Eqs. (13) are solved under the conditions, Eqs. (15). This done, the coordinates x_n of the particles and the pressures and viscous forces between neighboring particles are given at any time t by means of Eqs. (11) and (12).

Another problem which may be treated by the same method is that in which the gas in a stationary closed tube, say, a shock tube, is divided into two parts by a membrane, one part being at pressure p_0, and the other at pressure $p_1=\alpha p_0$. Initially the gas is at rest with constant density ρ_0 when the membrane is removed. The corresponding conditions for the model are in this

case

$$\begin{aligned} &T=0: \quad X_n=\frac{n}{N+1}, \quad X_n'=0; && n=1,\cdots,N \\ &\qquad P_{n+\frac{1}{2}}=\alpha; && n=0,\cdots,N_0-1 \\ &\qquad P_{n+\frac{1}{2}}=1; && n=N_0,\cdots,N \\ &T\geq 0: \quad X_0=0, \qquad X_n=\frac{n}{N+1}; && n=N_1,\cdots,N+1.\dagger \end{aligned} \tag{16}$$

Here N_0 is the particle nearest the membrane on the left and $N_1\leq N+1$. In this problem the dependence of the initial conditions on a parameter is removed and Eqs. (13) have a one-parameter family of solutions when α and N_0 are specified. It will therefore be noted from Eqs. (11) and (12) that since V disappears from the transformations to dimensional quantities, as well as from the parameter A, see Eq. (14), its definition is not necessary for this problem.

FINITE-DIFFERENCE REPRESENTATION

In order to be able to carry out the solution of Eqs. (13) on a digital calculator, it is further necessary to replace these by an equivalent system in finite-difference form. This can be accomplished in many different ways. Special care, however, must be taken to insure that the finite-difference representation chosen is numerically stable in the sense discussed in references 1 and 2. With this requirement in mind, the following system of equations was selected for numerical calculations:

$$\begin{cases} \Sigma_{n+\frac{1}{2}}^{m+\frac{1}{2}}=\dfrac{2}{\Delta T}\left[\dfrac{(X_{n+1}^{m+1}-X_{n+1}^{m})-(X_n^{m+1}-X_n^{m})}{(X_{n+1}^{m+1}+X_{n+1}^{m})-(X_n^{m+1}+X_n^{m})}\right]; \\ \qquad \Sigma_{n+\frac{1}{2}}^{m}=\tfrac{1}{2}[\Sigma_{n+\frac{1}{2}}^{m+\frac{1}{2}}+\Sigma_{n+\frac{1}{2}}^{m-\frac{1}{2}}] \\ X_n^{m+1}-2X_n^{m}+X_n^{m-1} \\ =(\Delta T)^2[-(P_{n+\frac{1}{2}}^{m}-P_{n-\frac{1}{2}}^{m})+A(\Sigma_{n+\frac{1}{2}}^{m}-\Sigma_{n-\frac{1}{2}}^{m})] \\ P_{n+\frac{1}{2}}^{m+1}-P_{n+\frac{1}{2}}^{m}=\Delta T\Sigma_{n+\frac{1}{2}}^{m+\frac{1}{2}} \\ \qquad \times\left[(\kappa-1)A\Sigma_{n+\frac{1}{2}}^{m+\frac{1}{2}}-\dfrac{\kappa}{2}(P_{n+\frac{1}{2}}^{m}+P_{n+\frac{1}{2}}^{m+1})\right], \end{cases} \tag{17}$$

where the notation $P_{n+\frac{1}{2}}^{m}$ signifies the pressure midway between the n and $(n+1)$th particles at time $T=m\Delta T$. Since $\Sigma_{n+\frac{1}{2}}^{m+\frac{1}{2}}$ is expressed implicitly in the system of Eqs. (17), the actual calculation was carried out by an iterative procedure, using the extrapolated value

$$\Sigma_{n+\frac{1}{2}}^{m}=\tfrac{3}{2}\Sigma_{n+\frac{1}{2}}^{m-\frac{1}{2}}-\tfrac{1}{2}\Sigma_{n+\frac{1}{2}}^{m-\frac{3}{2}}$$

as a first approximation.

An analysis of the stability of the system of Eqs. (17) was carried out. The results indicated that in regions where the viscous forces are negligible the solution is stable provided $\Delta T\leq(X_n-X_{n-1}/\kappa P)^{\frac{1}{2}}$. This is equivalent to the condition $\Delta T/\Delta X\leq 1/\text{local acoustic speed}$. In regions where the viscosity terms are predominant the solution is always stable. The stability argument is given in the Appendix.

SAMPLE CALCULATIONS

In order to illustrate the feasibility of carrying out the solution of one-dimensional flow problems involving shock waves by the proposed technique, three sample calculations were carried out on the Naval Ordnance Laboratory calculators. Both types of problems discussed earlier in the text were treated. More specifically, calculations were carried out for the following cases.

Case 1(a)

Type of problem: motion of a gas in a closed tube.
Number of particles: 99 moving.
Initial and boundary values:

$$\begin{aligned} &T=0: \quad X_n=\frac{n}{100}, \quad X_n'=\left(\frac{\kappa}{100}\right)^{\frac{1}{2}}M=0.0340; \\ &\qquad\qquad\qquad\qquad\qquad\qquad n=1,\cdots,99 \\ &\qquad P_{n+\frac{1}{2}}=1; \qquad\qquad\qquad\qquad n=0,\cdots,99 \\ &T\geq 0: \quad X_0=0, \qquad X_{100}=1. \end{aligned}$$

Constants:

$$A=10^{-4}, \quad \Delta T=0.01, \quad N+1=100.$$

These values of A and N cover the case $l=100$ cm and approximately standard initial conditions (and viscosity

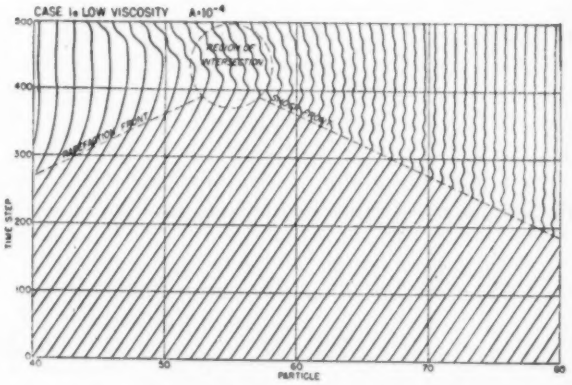


FIG. 1. World lines, Case 1(a) (low viscosity).

† It was found convenient in the numerical computations to take a model with $N+1=100$, although the particles $n=40,\cdots,100$ were kept fixed for all time.

[1] J. von Neumann and R. D. Richtmeyer, J. Appl. Phys. **21**, 232 (1950).

[2] O'Brien, Hyman, and Kaplan, J. Math. Phys. **29**, 223 (1951).

for air). The initial velocity is about one-third the initial acoustic speed.

Case 1(b)

Type of problem: motion of a gas in a closed tube.
Number of particles: 99 moving.
Initial and boundary values:

$$T=0: \quad X_n=\frac{n}{100}, \quad X_n'=\left(\frac{\kappa}{100}\right)^{\frac{1}{2}}M=0.0340; \qquad n=1, \cdots, 99$$

$$P_{n+\frac{1}{2}}=1; \qquad n=0, \cdots, 99$$

$$T\geq 0: \quad X_0=0, \qquad X_{100}=1.$$

Constants:

$$A=10^{-1}, \quad \Delta T=0.01, \quad N+1=100.$$

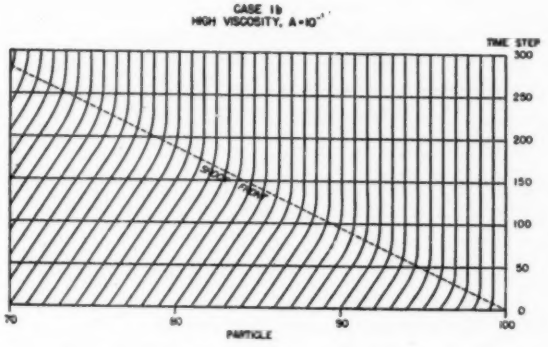


FIG. 2. World lines, Case 1(b) (high viscosity).

Case 2

Type of problem: shock tube.
Number of particles: 39 moving.
Initial and boundary values:

$$T=0: \quad X_n=\frac{n}{100}, \qquad X_n'=0; \qquad n=1, \cdots, 99$$

$$P_{n+\frac{1}{2}}=\alpha=5; \qquad n=0, \cdots, 19$$

$$P_{n+\frac{1}{2}}=1; \qquad n=20, \cdots, 99$$

$$T\geq 0: \quad X_0=0, \qquad X_n=\frac{n}{100}; \qquad n=40, \cdots, 100.$$

Constants:

$$A=10^{-1}, \quad \Delta T=0.01, \quad N_0=20, \quad N_1=40, \quad N+1=100.$$

Cases 1(a) and 1(b) represent models which simulate the motion of a gas confined in a tube closed at both ends with an initial velocity of 0.0340 (in the non-dimensional units). In case 1(a) a low viscosity coefficient is used which is of the order of magnitude expected in a real gas for a tube of length $l=100$ cm and standard conditions. Since the interval used in this calculation is larger by order of magnitude than the thickness of the shock front, the effects of viscosity are entirely lost, and the solution is characterized by oscillations (see Fig. 1) of the particles behind the "shock front," as has been found by von Neumann in reference 3. Thus, the solution describes accurately the motion of a mechanical model (of "beads" and springs[3]) rather than the motion of a gas in a tube. This can be corrected either by using a space (and time) interval of the order of magnitude of the shock wave thickness, or by introducing an unrealistically high value for the viscosity coefficient (or initial pressure) and thus increasing the shock wave thickness to the magnitude of the integration interval. Since it is practically unfeasible to carry out the former, a high value for the viscosity coefficient is used in the subsequent calculations [Cases 1(b) and 2]. This means that the calculations are being carried out for a gas with an unrealistically high viscosity and are not strictly correct for a gas such as air. However, since the effects caused by viscosity are considerable only within the shock front where velocity gradients are large, it is believed that in many respects the results obtained describe accurately the motion of a real gas.

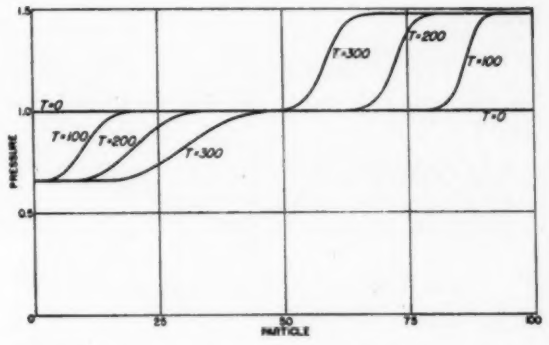


FIG. 3. Pressure *vs* particle number, Case 1(b), pressure *vs* particle number for typical time steps.

In Case 1(b) the conditions are identical with those used in Case 1(a) with the exception that the value for A is taken to be 10^{-1} rather than 10^{-4}. Figure 2 presents a plot of the world lines for this case. It will be noticed that the oscillations appearing in Fig. 1 [Case 1(a)] have now been eliminated. Figure 3 gives a plot of the pressures for Case 1(b) at $T=1.00$, 2.00, and 3.00 in non-dimensional units.

Case 2 represents a model which is used to simulate the flow in a shock tube. Forty particles were used in this calculation, and the calculation was carried to

[3] J. von Neumann, "Proposal and analysis of a new numerical method for the treatment of hydrodynamical shock problems," Applied Mathematics Panel, Appl. Math. Rept. 108.1R, March, 1944.

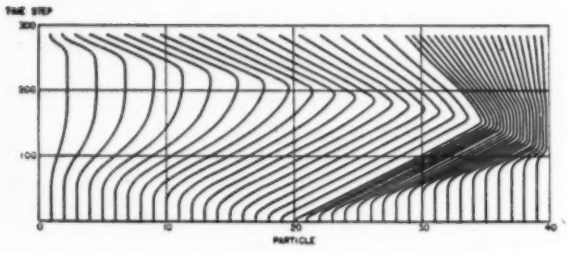


FIG. 4. World lines, Case 2 with pressure-temperature discontinuity.

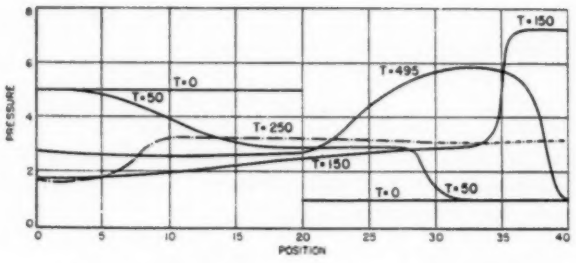


FIG. 5. Pressure *vs* particle number, Case 2 with pressure-temperature discontinuity.

$T=4.95$ (495 time steps). A high pressure ($P=5$, in the nondimensional units) and temperature was assumed to exist initially in one half of the shock tube and a low pressure (and temperature) in the remaining portion. For the purpose of this calculation the initial density was assumed constant throughout the shock tube. These are not the usual conditions which prevail initially in a shock tube. Ordinarily the pressure and density are increased at one end of a shock tube, whereas temperature is kept constant. Figure 4 presents a plot of the world lines for Case 2; and Fig. 5 gives a plot of the pressures for several time values.

The authors would like to express their appreciation to T. S. Walton, B. F. Cheydleur, R. P. Eddy, and R. Graffius for planning and carrying out the computations on the Naval Ordnance Laboratory calculators, and to H. H. Goldstine for advice in connection with the stability analysis.[4]

APPENDIX

Stability of Finite-Difference Equation System (17)

A stability analysis is carried out along the lines proposed by von Neumann as discussed in references 1 and 2. A system of difference equations is said to be stable provided the amplitude of a small disturbance, introduced at any time T, does not increase with successive time steps. More precisely, if $\delta F(X, T)$ and $\delta F(X, T+\Delta T)$ are the values of a variation of any of the dependent variables F in the system, then the system is said to be stable when $|\delta F(X, T+\Delta T)/\delta F(X, T)| \leq 1$. In solving systems of differential equations by finite-difference methods, extreme care must be taken to insure that the finite-difference representation is of this character, or the result may be completely invalidated. It will be shown here that the system of Eqs. (17) is essentially stable, provided the time interval ΔT is taken sufficiently small so that the inequality $\Delta T \leq (\Delta X/\kappa P)^{\frac{1}{2}}$ is satisfied. For the purpose of this analysis regions of negligible viscosity and regions in which the viscous terms are predominant (such as in the interior of the shock front) will be considered separately.

We will first assume that the viscosity terms in Eq. (17) may be neglected in comparison with pressure terms. We will introduce perturbations δX, δP, and $\delta\Sigma$ in each of the functions X, P, and Σ. Substituting in Eq. (17), we obtain the variational system of equations,

$$\begin{cases} \delta X_n^{m+1} - 2\delta X_n^m + \delta X_n^{m-1} + (\Delta T)^2[\delta P_{n+\frac{1}{2}}^m - \delta P_{n-\frac{1}{2}}^m] = 0, \\ \delta P_{n+\frac{1}{2}}^{m+1} - \delta P_{n+\frac{1}{2}}^m + \dfrac{\kappa\Delta T}{2}[\Sigma_{n+\frac{1}{2}}^{m+\frac{1}{2}}(\delta P_{n+\frac{1}{2}}^m + \delta P_{n+\frac{1}{2}}^{m+1}) + (P_{n+\frac{1}{2}}^m + P_{n+\frac{1}{2}}^{m+1})\delta\Sigma_{n+\frac{1}{2}}^{m+\frac{1}{2}}] = 0, \\ \dfrac{\Delta T}{2}\Sigma_{n+\frac{1}{2}}^{m+\frac{1}{2}}[(\delta X_{n+1}^{m+1} + \delta X_{n+1}^m) - (\delta X_n^{m+1} + \delta X_n^m)] + \dfrac{\Delta T}{2}[(X_{n+1}^{m+1} + X_{n+1}^m) - (X_n^{m+1} + X_n^m)]\delta\Sigma_{n+\frac{1}{2}}^{m+\frac{1}{2}} \\ \qquad - [(\delta X_{n+1}^{m-1} - \delta X_{n+1}^m) - (\delta X_n^{m+1} - \delta X_n^m)] = 0. \end{cases} \tag{18}$$

It will be further assumed that we are dealing with a small region in the (X, T) plane and that for such a region the coefficients ($\Sigma_{n+\frac{1}{2}}^{m+\frac{1}{2}}$, $P_{n+\frac{1}{2}}^m$, etc.) of the variational functions can be treated as constants. A solution of the system (18) can then be obtained in the form

$$\delta X_n^m = ae^{i\beta n + \alpha m\Delta T}; \quad \delta\Sigma_n^m = be^{i\beta n + \alpha m\Delta T}; \quad \delta P_n^m = ce^{i\beta n + \alpha m\Delta T}, \tag{19}$$

where a, b, c are real constants, α complex. Substituting Eq. (19) in Eqs. (18), we obtain a system of linear homogeneous equations for the quantities a, b, and c which has a nontrivial solution provided the determinant D of the

[4] H. H. Goldstine, private communication (1950).

coefficients is identically zero, where

$$D=\begin{vmatrix} \lambda-2+\frac{1}{\lambda} & 2i(\Delta T)^2\sin\frac{\beta}{2} & 0 \\ (e^{i\beta}-1)\left[\frac{\Delta T}{2}\Sigma_{n+\frac{1}{2}}{}^{m+\frac{1}{2}}(\lambda+1)-(\lambda-1)\right] & 0 & \frac{\Delta T}{2}e^{i\beta/2}\lambda^{\frac{1}{2}}[(X_{n+1}{}^{m+1}+X_{n+1}{}^{m})-(X_n{}^{m+1}+X_n{}^{m})] \\ 0 & (\lambda-1)+\frac{\kappa\Delta T}{2}\Sigma_{n+\frac{1}{2}}{}^{m+\frac{1}{2}}(\lambda+1) & \frac{\kappa\Delta T}{2}\lambda^{\frac{1}{2}}(P_{n+\frac{1}{2}}{}^{m+1}+P_{n+\frac{1}{2}}{}^{m}) \end{vmatrix}=0,$$

and where $\lambda=e^{\alpha\Delta T}$.

This equation is reduced to

$$(\lambda-1)[\lambda^2-2(1-L)\lambda+1]=0,$$

where

$$L=2\kappa\sin^2\frac{\beta}{2}(\Delta T)^2\frac{\langle P\rangle}{\langle\Delta X\rangle},$$

$$\langle P\rangle=\tfrac{1}{2}[P_{n+\frac{1}{2}}{}^{m+1}+P_{n+\frac{1}{2}}{}^{m}];\quad \langle\Delta X\rangle=\tfrac{1}{2}[(X_{n+1}{}^{m+1}+X_{n+1}{}^{m})-(X_n{}^{m+1}+X_n{}^{m})],$$

provided terms in $\Sigma\Delta T$ are neglected. It follows that $|\lambda|\leq 1$ for all β whenever

$$\Delta T\leq\left(\frac{\langle\Delta X\rangle}{\kappa\langle P\rangle}\right)^{\frac{1}{2}}.$$

Since $\langle\Delta X\rangle$ and $\langle P\rangle$ are average values for the difference in X and pressure, respectively, this result may be written simply $\Delta T\leq(\Delta X/\kappa P)^{\frac{1}{2}}$ or

$$\Delta T/\Delta X\leq 1/\text{local acoustic speed}.$$

On the other hand, if we assume that pressure terms may be neglected in comparison with the viscosity terms, we obtain the following system of variational equations:

$$\begin{aligned} &\delta X_n{}^{m+1}-2\delta X_n{}^{m}+\delta X_n{}^{m-1}-A(\Delta T)^2[\delta\Sigma_{n+\frac{1}{2}}{}^{m}-\delta\Sigma_{n-\frac{1}{2}}{}^{m}]=0,\\ &\delta P_{n+\frac{1}{2}}{}^{m+1}-\delta P_{n+\frac{1}{2}}{}^{m}-2\Delta TA(\kappa-1)\Sigma_{n+\frac{1}{2}}{}^{m+\frac{1}{2}}\delta\Sigma_{n+\frac{1}{2}}{}^{m+\frac{1}{2}}=0,\\ &\frac{\Delta T}{2}\Sigma_{n+\frac{1}{2}}{}^{m+\frac{1}{2}}[(\delta X_{n+1}{}^{m+1}+\delta X_{n+1}{}^{m})-(\delta X_n{}^{m+1}+\delta X_n{}^{m})]+\frac{\Delta T}{2}[(X_{n+1}{}^{m+1}+X_{n+1}{}^{m})-(X_n{}^{m+1}+X_n{}^{m})]\delta\Sigma_{n+\frac{1}{2}}{}^{m+\frac{1}{2}}\\ &\qquad-[(\delta X_{n+1}{}^{m+1}-\delta X_{n+1}{}^{m})-(\delta X_n{}^{m+1}-\delta X_n{}^{m})]=0, \end{aligned}\tag{20}$$

where

$$\delta\Sigma_{n+\frac{1}{2}}{}^{m}=\tfrac{1}{2}[\delta\Sigma_{n+\frac{1}{2}}{}^{m-\frac{1}{2}}+\delta\Sigma_{n+\frac{1}{2}}{}^{m+\frac{1}{2}}].$$

Substituting Eqs. (19) in (20) and equating the determinant of the coefficients of a, b, and c to zero, we obtain

$$D=\begin{vmatrix} \lambda-2+\frac{1}{\lambda} & 0 & -A(\Delta T)i\sin\frac{\beta}{2}(\lambda^{\frac{1}{2}}+\lambda^{-\frac{1}{2}}) \\ 0 & (\lambda-1) & -2\Delta TA(\kappa-1)\lambda^{\frac{1}{2}}\Sigma_{n+\frac{1}{2}}{}^{m+\frac{1}{2}} \\ 2i\sin\frac{\beta}{2}\left[\frac{\Delta T}{2}\Sigma_{n+\frac{1}{2}}{}^{m+\frac{1}{2}}(\lambda+1)-(\lambda-1)\right] & 0 & \lambda^{\frac{1}{2}}\frac{\Delta T}{2}[(X_{n+1}{}^{m+1}+X_{n+1}{}^{m})-(X_n{}^{m+1}+X_n{}^{m})] \end{vmatrix}=0.$$

Neglecting terms in $\Delta T\Sigma$, this reduces to

$$(\lambda-1)[(\lambda-1)^2[(X_{n+1}{}^{m+1}+X_{n+1}{}^{m})-(X_n{}^{m+1}+X_n{}^{m})]+4(\lambda^2-1)\sin^2(\beta/2)]=0.$$

Since $[(X_{n+1}{}^{m+1}+X_{n+1}{}^{m})-(X_n{}^{m+1}+X_n{}^{m})]>0$ it follows that $|\lambda|\leq 1$ for all β and all ΔT.

In conclusion, it is pointed out that in establishing the stability of Eqs. (17) for $\Delta T\leq(\Delta X/\kappa P)^{\frac{1}{2}}$, certain assumptions and restrictions were made. This analysis, therefore, does not insure the stability of the numerical procedure imposed under all circumstances, but is subject to the restrictions made. It should be noted, however, that the calculations carried out at the Naval Ordnance Laboratory based on the above criterion have shown no signs of instability, and thus offer evidence that the conditions imposed in the analysis are valid.

Letters to the Editor

A Simple Method for Measuring Interface Impedance*

A. VAN DER ZIEL
Electrical Engineering Department, University of Minnesota, Minneapolis, Minnesota
(Received January 23, 1953)

IN some vacuum tubes a high resistance interface layer develops between the nickel and the coating if the tube is aged. The equivalent circuit of this layer consists approximately of a resistance R_c and a capacity C_c in parallel, so that the impedance Z_c of the layer is

$$Z_c=R_c/(1+j\omega C_cR_c). \quad (1)$$

We propose here a simple method for determining R_c and C_c, which consists of an adaptation of a method used by Strutt and van der Ziel[1] for the measurement of the (complex) transconductance of vhf vacuum tubes. In that method a variable impedance Z is inserted between the grid and the anode of the tube under investigation; the tube is connected as an amplifier, and the impedance Z is adjusted such that the amplifier has zero output; the "apparent" transconductance g_m' of the tube (that is, the transconductance which is actually measured) is then

$$g_m'=1/Z \quad \text{or} \quad 1/g_m'=Z. \quad (2)$$

If the variable impedance Z consists of a variable resistance R_1 in series with a variable resistance R_2 and a variable capacity C_2 in parallel connection, then the output will be zero if (Fig. 1)

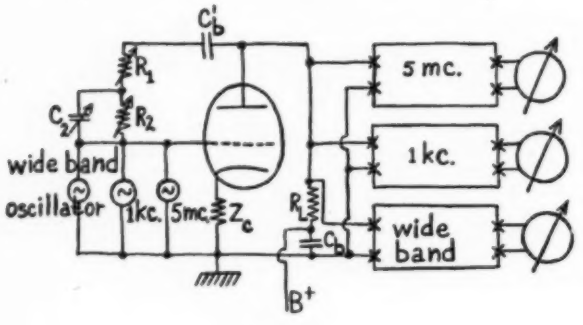


FIG. 1.

$$1/g_m'=R_1+R_2/(1+j\omega C_2R_2). \quad (3)$$

Now it may be shown that the impedance Z_c gives rise to an "apparent" transconductance,

$$g_m'=g_m/(1+g_mZ_c) \quad \text{or} \quad 1/g_m'=1/g_m+R_c/(1+j\omega C_cR_c), \quad (4)$$

as long as $|Z_c|\ll r_p$, where g_m is the *true* transconductance and r_p the plate resistance of the tube.

Comparing Eqs. (4) and (3), we see that R_1, R_2, and C_2 may be adjusted such that the output of the amplifier is zero at all frequencies. Apparently this is the case if

$$1/g_m=R_1; \quad R_c=R_2; \quad \text{and} \quad C_c=C_2. \quad (5)$$

In the actual measurement three oscillators should be connected to the input and three amplifiers should be connected across the load resistance R_L. The one oscillator should supply a frequency such that $Z_c\simeq 0$; in the figure we have chosen a frequency of 5 Mc, but that value might profitably been chosen somewhat smaller. A corresponding amplifier should be tuned to that frequency. The second oscillator should supply a frequency such that $Z_c\simeq R_c$, (in the figure we have chosen a frequency of 1 kc); a corresponding amplifier should be tuned to that frequency. A third oscillator should be a wide band oscillator, e.g., covering the range between 50 kc and 1000 kc; the output for that frequency should be measured with the help of a wide band receiver, which does not respond to the other two oscillators. The blocking condenser C_b' should be chosen such that it has negligible impedance at 1 kc; the load resistance R_L in the anode lead should not be chosen too large.

In the measurement we start with $R_2=0$ and adjust R_1 such that the 5-Mc channel gives no output; we then have

$$1/g_m=R_1.$$

We then adjust R_2 such that the 1-kc channel gives zero output; then

$$R_c=R_2.$$

We finally set the wide band oscillator at such a frequency that the wide band amplifier gives a reasonable output. We then adjust C_2 such that the wide band channel gives zero output. This indicates that

$$\omega C_cR_c=\omega C_2R_2 \quad \text{or} \quad C_c=C_2.$$

The three tube parameters g_m, R_c, and C_c are thus determined by null methods.

If the signal level is kept low, it is also possible to use the instrument for a study of the nonlinearity of the interface layer.

* Work performed under U. S. Signal Corps Contract.
[1] M. J. O. Strutt and A. van der Ziel, Elek Nachr. Tech. 15, 103 (1938).

The Potential of Two Current Point Sources in a Homogeneous Conducting Prolate Spheroid

JAMES R. WAIT
Radio Physics Laboratory, Defense Research Board, Ottawa, Canada
(Received December 15, 1952)

IN the theoretical aspects of electrocardiography, the human body is often assumed to be a homogeneous, isotropic, and conducting mass. For convenience its shape was assumed to be spherical by Frank[1] in a recent interesting paper. He has developed expressions for the resulting potentials when current point electrodes are embedded in the sphere. A better model might be the conducting prolate spheroid that has point current electrodes at any specified location in the interior or on the surface.

Since the solution of this problem may be of interest, it is given here. A prolate spheroidal body of homogeneous conductivity σ has a major axis b and semiminor axis a. The equation of its surface is given by

$$\rho^2/a^2+z^2/b^2=1, \quad (1)$$

where a conventional cylindrical coordinate system (ρ, ϕ, z) is chosen so that the spheriod is coaxial with the z axis. A current point source supplying a current I is located at (ρ_1, ϕ_1, z_1) inside the spheroid. A current point sink carries the return current I and is located at (ρ_2, ϕ_2, z_2), which is also inside the spheroid.

It is now convenient to introduce prolate spheroidal coordinates (η, ϕ, δ). The radial coordinate η specifies a family of spheroids given by

$$\frac{\rho^2}{c^2(\eta^2-1)}+\frac{z^2}{c^2\eta^2}=1, \quad (2)$$

and the angular coordinate δ specifies a family of hyperboliods given by

$$\frac{\rho^2}{-c^2(1-\delta^2)}+\frac{z^2}{c^2\delta^2}=1. \quad (3)$$

The constant c is chosen so that family of spheroids in Eq. (2) is confocal with the prolate spheroidal body, and it is therefore given by

$$c=(b^2-a^2)^{\frac{1}{2}}.$$

The equations of transformation are

$$z=c\eta\delta, \qquad \rho=c[(1-\delta^2)(\eta^2-1)]^{\frac{1}{2}}. \quad (4)$$

In this coordinate system the surface of the prolate spheroidal body is specified by $\eta=\eta_0$, and the current point source C_1 and sink C_2 are located at $(\eta_1, \phi_1, \delta_1)$ and $(\eta_2, \phi_2, \delta_2)$, respectively.

The potential V_1 due to the source C_1 and the sink C_2, at a point (η, ϕ, δ) inside the spheriod, is a solution of Laplaces equations, so it can be represented in terms of spheroidal harmonics[2] as follows:

$$V=\sum_{n=0}^{\infty}\sum_{m=0}^{n}[A_{mn}Q_n^m(\eta)+B_{mn}P_n^m(\eta)]P_n^m(\delta) \quad (5)$$

for the region $\eta_0\geq\eta\geq\eta_1$ and η_2. Now if the spheroid were of infinite dimensions such that $\eta_0\to\infty$, the potential V_1 becomes

$$V=\frac{I}{4\pi\sigma}\left(\frac{1}{R_1}-\frac{1}{R_2}\right)=\sum_{n=0}^{\infty}\sum_{m=0}^{n}A_{mn}Q_n^m(\eta)P_n^m(\delta), \quad (6)$$

where R_1 is the distance between the points (η, ϕ, δ) and $(\eta_1, \phi_1, \delta_1)$ and R_2 is the distance between the points (η, ϕ, δ) and $(\eta_2, \phi_2, \delta_2)$. By making use of a well-known addition formula for Legendre functions[3] the coefficient A_{mn} is given by

$$A_{m0}=0,$$

and for $n>0$

$$A_{mn}=\frac{I}{4\pi\sigma c}\epsilon_m(-1)^m(2n+1)\frac{(n-m)!}{(n+m)!}\times\sum_{i=1,2}(-1)^{i-1}P_n^m(\eta_i)P_n^m(\delta_i)\cos m(\phi-\phi_i), \quad (7)$$

where

$$\epsilon_m=1 \quad \text{if} \quad m=0$$
$$=2 \quad \text{if} \quad m\neq 0.$$

The coefficient B_{mn} can now be found from the condition that the normal flow or the normal derivative of the potential shall be zero at the surface $\eta=\eta_0$. This gives rise to the following relation:

$$A_{mn}Q_n^{m\prime}(\eta_0)+B_{mn}P_n^{m\prime}(\eta_0)=0, \quad (8)$$

where the prime indicates a differentiation with respect to η_0.

The potential V at (η, ϕ, δ), for $\eta\leq\eta_0$, due to both source and sink is then explicitly given by

$$V=\frac{I}{4\pi\sigma}\sum_{i=1,2}(-1)^{i-1}\left\{\frac{1}{R_i}-\frac{1}{c}\sum_{n=1}^{\infty}\sum_{m=0}^{n}\epsilon_m(-1)^m(2n+1)\frac{(n-m)!}{(n+m)!}\times\frac{Q_n^{m\prime}(\eta_0)}{P_n^{m\prime}(\eta_0)}P_n^m(\delta)P_n^m(\eta)P_n^m(\delta_i)P_n^m(\eta_i)\cos m(\phi-\phi_i)\right\}. \quad (9)$$

The associated Legendre functions in the preceding expression are extensively tabulated so that numerical values for any interesting cases are computable.

[1] Ernest Frank, J. Appl. Phys. 23, 1225 (1952).
[2] E. W. Hobson, *Spherical and Ellipsoidal Harmonics* (Cambridge University Press, Cambridge, 1931).
[3] E. Jahnke and F. Emde, *Tables of Functions* (Dover Publications, New York, 1945), p. 115.

A Method for Increasing the Electrical Resistivity of Insulators Under Ionizing Radiation

JOHN H. COLEMAN AND DAVID BOHM *
Radiation Research Corporation, West Palm Beach, Florida
(Received January 26, 1953)

IT has been known for some time that the volume resistivity of good insulators is lowered by ionizing radiation. Farmer[1] reported a decrease in resistivity of several decades for polystyrene under x-radiation. United States Atomic Energy Commission reports[2-3] covered the reduction in resistivity in insulators under β- and γ-radiation. These measurements, however, were made after the radiation was removed. Measurements taken in this laboratory on good insulators under β-radiation indicated the induced specific volume resistivity decreased for a definite period of time, reached a minimum, and then actually increased until physical degradation resulted in surface leakage.

Figure 1 shows this initial decrease in a 0.002-inch thick polysturene sheet which was measured in the form of a parallel plane condenser and irradiated with a 25-mc Sr90 β-source deposited on one electrode. The other electrode collected the B-particles that penetrated the insulator. A 1500-volt battery and electrometer were electrically connected between the two electrodes, and the resulting current through the insulator was recorded. Figure 2

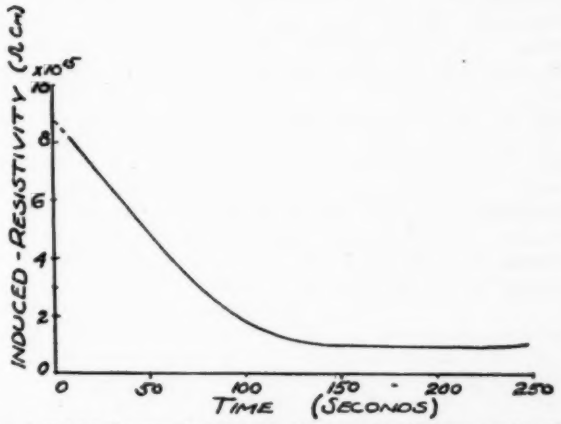


FIG. 1. Initial decrease in induced resistance of polystyrene under 25-mc strontium 90 radiation (β-current density = 12 micromicroamp per cm^2).

shows a log-log plot of the subsequent increase in resistivity after a broad minimum of approximately one day. The resistivity is seen to increase as the square root of the time.

This resistivity was found to be constant for various voltages and reversals of electric field after the usual polarization current decayed. These tests indicate this increase in radiation resistance is not determined by space charge which causes the well-known decrease in counting efficiency of crystal counters described by Hofstader.[4]

The radiation was removed after about one month, and the resistivity returned to its previous unirradiated value ($\rho>10^{18}$ ohm cm). On reappliaction of the radiation, however, the induced resistivity returned to the value measured just before the radiation

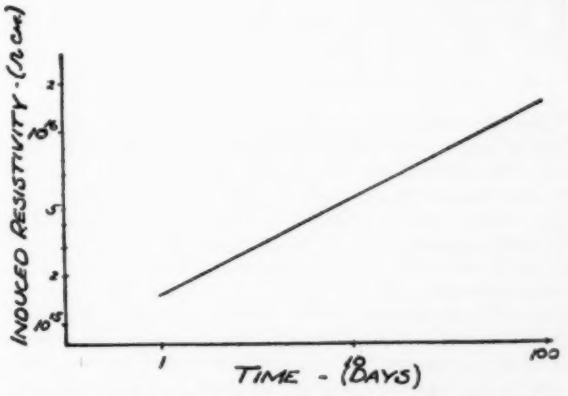


FIG. 2. Increase in induced resistance (continued irradiation of polystyrene shown in Fig. 1).

was removed, instead of going through the minimum again. Furthermore, baking the sample at 65°C after irradiation did not alter the conductivity on reapplication of the radiation. This "memory" of radiation resistance, however, could be partially erased by application of both heat and electric field.

This effect was found in all "good" insulators tested (i.e., unirradiated $\rho>10^{16}$ ohm cm). Typical minimum induced resis-

tivities found were polystyrene, 10^{15}; polyethylene, 5×10^{14}; polymonochlorotrifluoroethylene, 5×10^{16}; polyester "Mylar," 10^{16}; amber, 6×10^{16}; and mica, 5×10^{16}, all in ohm cm. The time to reach these values in every case was less than 30 minutes of irradiation. All materials but polystyrene showed some physical degradation.

For a theoretical description of the induced resistivity, Bohm[5] proposed a theory based on trapping[6] of electrons and holes which states, qualitatively, that the initial decrease in resistance is due to the charge carriers being released from traps after about 30 seconds. These carriers are produced by the radiation and originally trapped in a time less than one second, giving an observable resistivity extrapolated at zero time on the graph. Using this picture, an equilibrium charge due to recombination will be reached to give a minimum value of resistivity. One possibility for the increase in resistivity is the formation of new traps by the radiation. A more complete theory will be the basis of a future paper.

These results indicate that high insulation resistance can be maintained under radiation if the insulator has been previously irradiated with a sufficient dose. This method of providing good insulation resistance under radiation has application in any electrical equipment subject to light ionizing radiation, since the induced resistivity effect should be the same. For example, insulation for γ-ray ionization chambers, x-ray equipment, and nuclear reactors could be irradiated prior to use by a strong β- or γ-source.

Tests are being continued on these effects under Signal Corps Contract No. DA-36-039 SC 42654 for application in radioactive batteries.

* Now at Universidade de Sao Paulo, Sao Paulo, Brazil, South America.
[1] F. T. Farmer, Nature **150**, 521 (1942).
[2] J. G. Burr and W. M. Garrison, U. S. Atomic Energy Commission Report No. AECD-2078.
[3] O. Sisman and C. D. Bopp, Oak Ridge National Laboratory Report No. ORNL-928.
[4] R. Hofstader, Nucleonics **4**, April (1949).
[5] D. Bohm, RRC Technical Report No. 3, July 15–31 (1951).
[6] G. L. Perason and J. Bardeen, Phys. Rev. **75**, 865 (1949).

Thermoelectric Effects in Molybdenum Disilicide

MARTIN J. ARVIN
Southern Illinois University, Carbondale, Illinois
(Received January 16, 1953)

THE thermal emf of molybdenum disilicide ($MoSi_2$) against platinum has been measured for the temperature range -60°C to 600°C. The thermal emf in microvolts is given by

$$E=5.13t+\tfrac{1}{2}(3.76)10^{-2}t^2-\tfrac{1}{3}(15.65)10^{-6}t^3.$$

The values are close to those for copper *vs* platinum, which might be taken as additional evidence of the metallic nature[1] of $MoSi_2$. No difference in thermal emf was found between samples from two sources. In one case,[2] the samples were $2.5\times1\times0.5$ cm in size; in the other,[3] a sample $6.5\times0.5\times0.5$ cm was used. The alloys were prepared by direct synthesis of Mo and Si powders of high purity mixed in stoichiometric proportions.

Iron-constantan thermocouples for measuring junction temperatures and platinum lead wires were silver soldered together to the ends of the sample. One end was held in contact with a metal rod heated electrically and the other end was cooled by ice water, or by dry ice and acetone. Junction temperatures were read from a portable potentiometer and simultaneously thermal emfs were read from an *L* and *N* type *K* potentiometer.

A freshly polished surface of $MoSi_2$ was explored according to the method described by Granville and Hogarth[4] using a tungsten wire probe electropolished to a diameter of 0.0004 inch. No electrical effects different from those for an ordinary metal appeared.

[1] F. W. Glaser, J. Appl. Phys. **22**, 103 (1951).
[2] American Electro Metal Corporation, Yonkers, New York.
[3] Fansteel Metallurgical Corporation, North Chicago, Illinois.
[4] J. W. Granville and C. A. Hogarth, Proc. Phys. Soc. (London) **B64**, 488 (1951).

Permanent Magnet Properties of Cobalt-Platinum Alloys

D. L. MARTIN AND A. H. GEISLER
General Electric Research Laboratory, The Knolls, Schenectady, New York
(Received January 26, 1953)

ADAMS, Hubbard, and Syeles in their paper entitled "A New Permanent Magnet from Powdered Manganese Bismuthide" [J. Appl. Phys. **23**, 1207 (1952)] claim that the coercive force (H_c) of Bismanol is not exceeded by any other magnet. Their conclusion is based on a comparison with the magnetic values of Neumann[1] for cobalt-platinum, one of the strongest permanent magnet materials known. We would like to point out that Neumann's values are not the best for comparison and that in recent work on cobalt-platinum alloys higher magnetic values have been obtained.[2,3] These recent results are compared with the authors' Bismanol results in Table I. The values for the platinum alloys compare very favorably with those of Bismanol.

Examination of the data for the cobalt-platinum alloys shows that the peak values of B_r, H_c, H_{ci}, and $(BH)_{max}$ are dependent upon the composition, the cooling rate from the disordering treatment, and the aging time and temperature. By selection of these factors a wide range of values can be obtained as illustrated by Table I. Unfortunately, the high cost of platinum rules out extensive use of the cobalt-platinum alloy for all except special applications where its high coercive force combined with ductility might offer some advantage over other materials.

TABLE I. Comparison of the magnetic properties of bismanol and cobalt-platinum.

Magnet	Peak B_i gauss	B_r gauss	H_c oersteds	H_{ci} oersteds	$(BH)_{max}$ million	Treatment
Bismanol	...	4300	3400	7000	4.3	
42 At%Co-Pt	6400	5700	3900	5000	7.2	1000°C, air cooled, 10 hr 600°C
42 At%Co-Pt	6000	5100	4400	6300	6.0	1000°C, air cooled, 30 hr 600°C
42 At%Co-Pt	5500	4800	4000	7000	5.1	1000°C, air cooled, 100 hr 600°C
48 At%Co-Pt	6500	5700	3500	5000	6.2	1000°C water quenched, 10 hr 600°C
48 At%Co-Pt	5700	4500	3600	5700	5.0	1000°C water quenched, 25 hr 600°C
48 At%Co-Pt	7000	6200	4100	5000	8.2	1000°C air cooled, 6 hr 600°C
48 At%Co-Pt	6400	5400	4700	6500	6.7	1000°C air cooled, 15 hr 600°C
48 At%Co-Pt	7100	6300	4100	4900	9.0	special cool,[a] 5 hr 600°C
48 At%Co-Pt	6300	5400	4500	6600	6.0	special cool,[a] 20 hr 600°C

[a] Cooled from 1000°C to 200°C in 10 minutes.

[1] H. Neumann, Arch. tech. Messen **69**, 38 (1937).
[2] Newkirk, Geisler, Martin, and Smoluchowski, J. Metals **188**, 1249 (1950).
[3] D. L. Martin and A. H. Geisler, U. S. Patent 2,622,050.

Comments on "A Note on Critical Reflections of Elastic Waves at Free Surfaces"

R. D. MINDLIN AND T. KANE
Department of Civil Engineering, Columbia University, New York, New York
(Received January 13, 1953)

IN a paper under the above title,[1] Goodier and Bishop resolved a difficulty which had been a puzzling one for a long time. Recently W. S. Jardetzky[2] criticized their result as being obtained by "rather artificial assumptions." In a reply,[3] Goodier restated his point of view in greater detail, although, it seems to us, the matter was explained clearly enough in the paper itself.

We are in full agreement with both the Goodier and Bishop paper and Goodier's reply. The only purpose of this letter is to

comment on a question, raised by Jardetzky, which Goodier did not answer.

The waves found by Goodier and Bishop contain a component whose amplitude increases, without limit, with increasing distance from the free surface. Regarding this component, Jardetzky remarks: "Whether a physical meaning can be attached to it or not, remains an open question." The answer to the question is, in fact, given by Goodier and Bishop, although it is somewhat hidden behind the cryptic phrase "with suitable development of the exponentials" preceding their Eq. (7). Omission of the higher order terms, in the development of the exponentials, requires that the product of the wave number and the distance from the free surface remain finite. Accordingly, their result describes the behavior of the waves in the neighborhood of the free surface, and it is inappropriate to raise questions regarding the behavior of the solution at infinity. Goodier and Bishop have shown, simply, that reflected plane waves, at grazing incidence, must contain a component whose amplitude, near the surface, increases linearly (to the first order) with departure from the surface. There is an infinite variety of such waves, each differing from the others in its variation of amplitude with depth; but all of them must have, to the first order, the linearly varying component in the neighborhood of the boundary. This situation is implied by Goodier and Bishop in their reference to the paper by Sauter.

We have recently completed a study of the reflection of straight-crested flexural waves at the edge of a plate. The usual difficulty with the limiting case of grazing incidence appeared, but it was clearly resolved by an application of the technique introduced by Goodier and Bishop.

[1] J. N. Goodier and R. E. D. Bishop, J. Appl. Phys. 23, 124 (1952).
[2] W. S. Jardetzky, J. Appl. Phys. 23, 1279 (1952).
[3] J. N. Goodier, J. Appl. Phys. 23, 1280 (1952).

Comment on the Preceding Letter of R. D. Mindlin and T. Kane

W. S. Jardetzky
Columbia University, New York, New York
(Received February 2, 1953)

IT was shown in my "Remark" (reference 2, preceding letter) that the final result of Goodier and Bishop (reference 1, preceding letter) is obtained when one takes care of a rule concerning the integrals of differential equations. The principal fact is that these integrals have to be taken in different forms for the nongrazing and grazing incidence, respectively. This mathematical evidence is not taken into account in Goodier's comment (reference 3, preceding letter) or in the letter of R. D. Mindlin and T. Kane, in which, on the contrary, the following assumption is made: "Recently, W. S. Jardetzky criticized their result . . ." (i.e., that of Goodier and Bishop).

It is possible that some investigators will prefer to understand the phenomenon under discussion on using other kind of reasoning. I have nothing to change, however, in the approach to the problem given in my "Remark" (reference 2, preceding letter).

Diffusion of D_2 from D_2O through Steel

FRANCIS J. NORTON
General Electric Research Laboratory, Schenectady, New York
(Received January 23, 1953)

HYDROGEN will permeate steel at room temperature when corrosion occurs, owing to the presence of air and liquid water on the steel surface.[1,2] An experiment was carried out to show unambigously that the source of the hydrogen is the water participating in the initial corrosion reaction, $Fe+H_2O=FeO+H_2$ rather than hydrogen already present in the steel.

Heavy water, D_2O, and air were introduced into an outer glass vessel at 90°C. The presence of a little liquid D_2O at the bottom insured saturation conditions. This outer vessel surrounded a steel vacuum chamber, which was kept very slightly cooler. Thus, liquid condensation and corrosion occurred on the exterior of this low carbon steel chamber, 200 cm^2 in wall area and 0.16 mm thick. The interior led directly to a mass spectrometer.

When the D_2O was introduced on the outside of the steel vacuum chamber, at 90°C, the D_2 peak at mass 4 appeared within three minutes, and a flow of 4×10^{-6} cm^3 per min (NTP) was extablished for D_2 through the steel wall. No D_2O peaks were found, indicating no leak in the steel chamber. When air was removed from the outside, the D_2O being left, the D_2 rate of permeation was greatly reduced, showing that oxygen from the air is involved in the process.

[1] W. Dällenbach, Electrotech. Z. 55, 85 (1934).
[2] F. J. Norton, J. Appl. Phys. 11, 262 (1940).

A Solubility Determination of Some Elemental Gases in Pyrex at 1170°C

JOHN J. NAUGHTON *
Frick Chemical Laboratory, Princeton University, Princeton, New Jersey
(Received February 2, 1953)

WHILE attempting a measurement at high temperatures, of the solubility of some elemental gases in silicate systems of geochemical interest, a solubility was determined for hydrogen in molten Pyrex that was considerably higher than the value that would be expected from previous work on similar systems.[1] Simultaneous attempts to measure gas solubility for silicate systems that contained large amounts of alkali oxide were rendered difficult, owing to the evaporation of some component, presumably the alkali. In test work with Pyrex (Corning No. 774) no such difficulty was encountered, and it was found feasible to extend measurements to a comparatively high temperature.

The apparatus used was the same as the type of equipment that is generally employed in the measurement of the adsorption of gases by powdered materials.[2] However, the adsorption tube was replaced by a silica bulb with the sample contained therein in a platinum crucible. A Nichrome wound furnace was used for temperatures up to 900°C. For higher temperatures an induction furnace was employed with a split platinum radiation shield around the platinum crucible and with water cooling of the silica bulb to prevent diffusion and escape of helium and hydrogen. The temperature of the crucible and contents was measured by a platinum, platinum-rhodium thermocouple welded directly to the crucible. Samples were outgassed in a separate high vacuum system at approximately 200°C above the temperature of the solubility determination; intermittent heating for as long as 26 hours was necessary in some cases.

The gases used were oxygen, argon, helium, and hydrogen. The measurement of the gas solubility was made by running a blank at temperature with the gas in question and with an empty crucible in the system. The amount of gas in the furnace bulb was determined for various pressures. The pressure was varied by introducing known amounts of gas from a gas pipet and was measured by means of a U-type gauge with a microscope cathetometer capable of being read to 0.001 mm. The crucible with the degassed sample was next sealed into the system. For a given temperature and pressure the amount of gas in the furnace bulb was measured at intervals until equilibrium was attained, a matter of from fifteen to thirty-five minutes. This process was repeated at other pressures for each temperature. Repeat runs were made in each case. The measured gas content of the furnace was corrected for the amount of gas in the furnace dead space as determined by the blank and for any volume change within the furnace bulb with and without the sample. Measurements were limited by the range of the cathetometer to a maximum pressure of a little over 10 mm. Because of the time consuming nature of the procedures, measurements in the high temperature range for Pyrex have been completed thus far at only one temperature (1170°C).

In the one other system studied, sodium disilicate, no measur-

able solubility was found for any of the gases used up to 800°C. Above this temperature gas diffusion and sample evaporation became troublesome for external heating. A spark discharge type of induction heater was adapted for use and determinations were made with Pyrex as test substance at 1170°C. No measurable solubility was found for oxygen, argon, or helium, but for hydrogen a solubility of 0.060 ± 0.004 ml (STP) per gram of sample was measured at 10 mm pressure. The solution process was found to be reversible. A slight increase of solubility with increase of pressure was evident, although the limited range studied makes interpretation impossible. At present an attempt is being made to extend this work to other systems and to higher pressures.

* Present address: Chemistry Department, University of Hawaii, Honolulu, T. H.

[1] G. A. Williams and J. B. Ferguson, J. Am. Chem. Soc. 46, 635 (1924).
[2] W. E. Barr and V. J. Anhorn, *Scientific Glass Blowing* (Instruments Publishing Company, Pittsburgh, 1949).

JOURNAL OF APPLIED PHYSICS VOLUME 24, NUMBER 4 APRIL, 1953

Announcements

The following **meetings** will be held in France, at the **University of Paris (Sorbonne)** in the summer of 1953:

24th–26th June: A colloquium sponsored by the Rockefeller Foundation and organized by the French National Bureau of Scientific Research (CNRS) on: "Water molecules in crystals and solids studied by electromagnetic waves (from x-rays to wavelengths in the radiofrequency range)"
Active participation restricted to 20 invited persons.
Secretary: Dr. A. Kastler, Laboratoire de Physique, ENS, 24 rue Lhomond, Paris 5e.

29th June–3rd July: General International Meeting on Molecular Spectroscopy (from x-rays to wavelengths in radiofrequency range).
Secretary: Dr. J. Lecomte, Laboratoire des Recherches Physiques, Sorbonne, 1 rue Victor Cousin, Paris 5e.

6th–10th July: Symposium organized by the Société francaise de Chimie physique, on "Recent Advances in Studies of Molecular Structure."
Secretary: Dr. M. Magat, Laboratoire de Chimie physique, 11 rue Pierre Curie, Paris 5e.

The annual **Cornell Summer Laboratory Course in Techniques and Applications of the Electron Microscope** will be given this summer from June 15 to June 27, 1953, by the Laboratory of Electron Microscopy in the Department of Engineering Physics. The course under the direction of Dr. Benjamin M. Siegel will have Dr. James Hillier of the RCA Laboratories, Princeton, New Jersey, and Dr. C. E. Hall of M.I.T., Cambridge, Massachusetts, as guest lecturers this year.

The course is designed for those research workers, institutional and industrial, who have recently entered the field of electron microscopy or who are now planning to undertake research problems involving applications of this instrument. Further inquiries should be addressed to Dr. Benjamin M. Siegel, Department of Engineering Physics, Rockefeller Hall Cornell University, Ithaca, New York.

Boston College has announced a special two weeks intensive **Course in Modern Industrial Spectrography** at Chestnut Hill, Massachusetts, from July 20 to July 31. The course is particularly designed for chemists and physicists from industries in the process of installing spectrographic equipment. Information on the course may be obtained from James J. Devlin, S.J., Physics Department, Boston College, Chestnut Hill 67, Massachusetts.

The United States Testing Company, Inc. announces the 16th **Summer Refresher Course in Textiles and Testing Techniques** (s-275) which will be conducted during the second half of the Teachers College Summer Session, Columbia University, July 27 to August 14, 1953. The course will be held at the Company's main laboratories in Hoboken, New Jersey.

National Electronics Conference. The ninth annual conference will be held September 28, 29, and 30, 1953, at the Hotel Sherman in Chicago. Dr. J. D. Ryder, head of the electrical engineering department, University of Illinois, Urbana, has been named president of the 1953 National Electronics Conference, Inc., and Dr. C. E. Barthel, Jr., Illinois Institute of Technology, Chicago was named chairman of the board.

Radio Corporation of America. Eighteen university students from fourteen different states have been awarded RCA Scholarships for the 1952–1953 academic year. Seven RCA Scholarships have been awarded for the first time at the University of Notre Dame, West Virginia University, University of Cincinnati, Franklin and Marshall College, University of Florida, Indiana University, and Wellesley College. Science Major at Wellesley, Miss Marijane Curran, is the first woman to receive an RCA Scholarship.